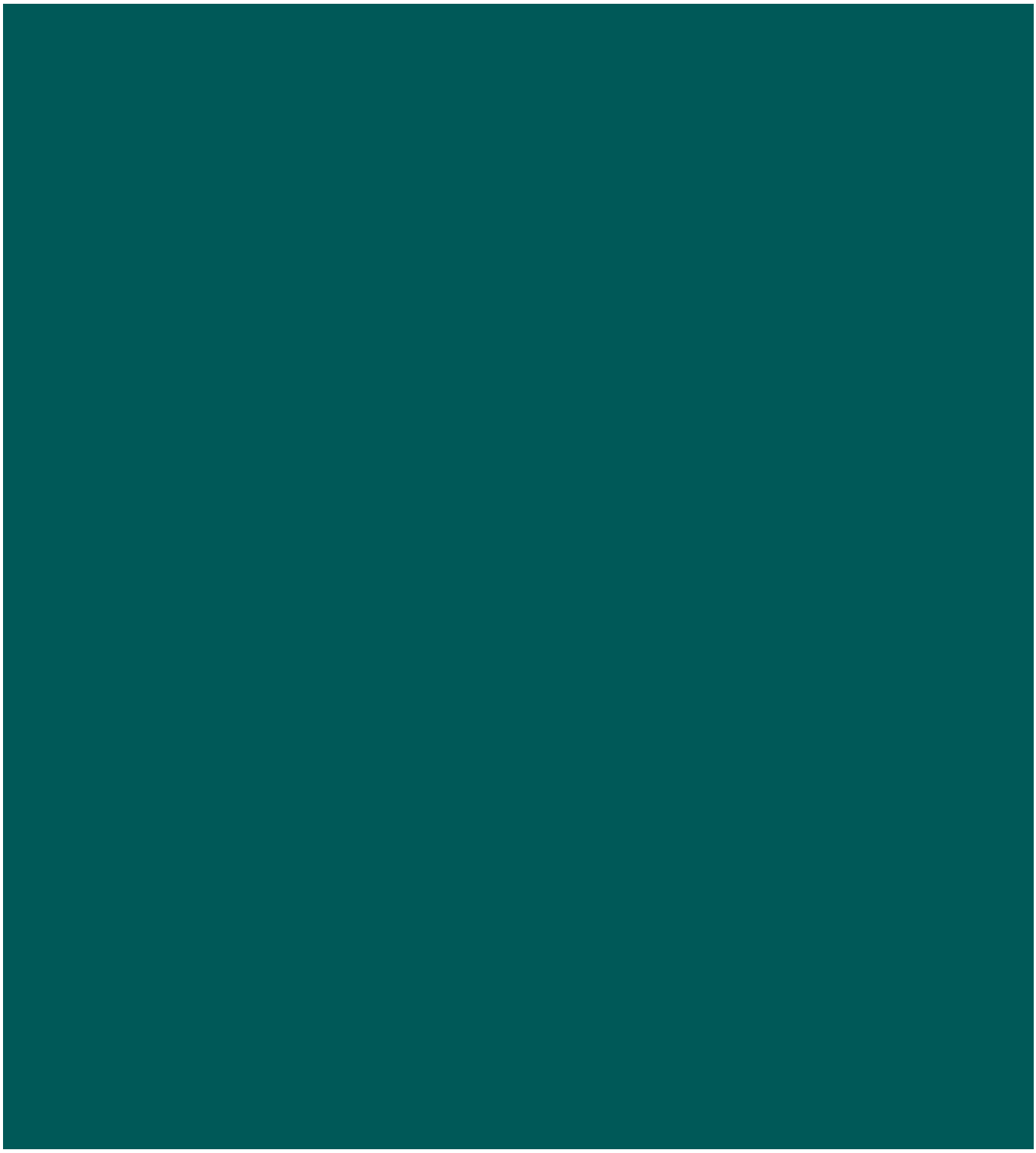
Boffa Miskell

Assessing the strength of scientific evidence for the development of science-informed policy

Prepared for the Ministry for the Environment

5 July 2024



Boffa Miskell is proudly a
Toitū net carbonzero certified consultancy

Document Quality Assurance

Bibliographic reference for citation: Boffa Miskell Ltd 2024. *Assessing the strength of scientific evidence for the development of science-informed policy.* Report prepared by Boffa Miskell Ltd for the Ministry for the Environment.		
Prepared by:	Bronte Linkhorn | Planner	
	Dr Tommaso Alestra | Ecologist	
	Dr Anne Cunningham | Engagement Specialist	
	Merryn Hedley | Librarian	
Reviewed by:	Deborah Rowe Planner | Senior Principal	
	Dr Tanya Blakely Ecologist | Associate Partner	
Status: FINAL	Revision / version: 3	Issue date: 5 July 2024

Template revision: 20230505 0000

File ref: BM240037_001b_Final Report_20240705_FINALdocx

Executive Summary

Introduction and context

It is widely accepted that scientific evidence should inform the development of effective policy. However, the process of developing evidence-based policy is a challenge experienced in all fields of policy making. Uncertainty, gaps and biases in understanding evidence can impact decision making and increase the risk of poor outcomes. Both the quality of evidence and / or the uptake of evidence can influence the quality of policy, guidelines and decision making.

Current barriers that may constrain the effective use of evidence in policy development include accessibility, relevance and applicability, organisational capacity, resources and finances, time constraints, and poor communication and dissemination skills between scientists and decision-makers (Cooke et al., 2023; S. J. Nichols et al., 2017).Some of the other key challenges that set the context for this issue include:

- Diminishing trust in political and scientific authorities
- Dealing with vested interests
- Increasing complexity of environmental problems
- Types of evidence

The policy cycle in New Zealand applies across various levels of government in New Zealand, including central, regional and local government. The ‘scale’ of policy-making processes can vary significantly, and in many instances associated processes may be described in legislation and may include statutory timeframes, which can place limitations on the time available to obtain, review, and synthesise evidence.

In light of the context and challenges outlined above, the Ministry for the Environment (the Ministry) has identified a need to ensure scientific evidence is assessed appropriately and informs the development of effective policy.

The objective of this project is to make recommendations that will inform the potential development of a repeatable, transparent process for assessing the strength of scientific evidence for the development of ‘science-informed policy’ that relates to environmental matters. The Ministry engaged Boffa Miskell to undertake this project.

Literature review methods

The scope of the literature review included identifying and evaluating the key principles, frameworks and methods used within New Zealand and internationally.

As the assessment of evidence in environmental policy is, internationally and within New Zealand, at an earlier stage of maturation compared to its application in other fields (e.g., education and health), there is also reference

made to frameworks and learnings from fields outside of environmental science.

A range of key search terms derived from the project objective were used to search databases (Google, Google Scholar) to gather initial literature material.

Reference lists from literature found were used to identify other relevant literature using similar search terms. A targeted review of overseas government websites was also carried out to obtain applied literature, such as guidelines for evidence-informed policy. Key themes were extracted, categorised, and analysed, then discussed among the Boffa Miskell researchers, reviewers, and a Challenge Group, to shape the literature review through an iterative process.

Assessing the strengths and weaknesses of the frameworks and methods

Strengths and weaknesses of each approach were assessed in relation to the following criteria established by the Ministry to identify an approach that:

- Allows for an ‘absolute’ assessment of evidence.
- Is transferable across policy questions in different environmental domains and at different levels of government.
- Is repeatable and allows for follow up assessments in the future.
- Can be applied to different stages of the policy cycle.
- Ensures a transparent evaluation of scientific evidence and reduces the likelihood of selection biases.

In considering the relative strengths and weaknesses of the various approaches in the literature, we have also been cognisant of some of the key challenges that can limit evidence-informed policy, such as those described by Cooke et al. (2023) and Nichols et al. (2017):

- accessibility
- relevance and applicability
- organisational capacity
- resources and finances
- time constraints
- poor communication and dissemination skills between scientists and decision makers.

This literature review focuses on scientific research evidence and does not include the assessment of other forms of evidence such as mātauranga Māori as it would be more appropriate to address this in a separate piece of work.

Use of frameworks

The literature consistently describes how introducing a framework guides the user through a set of structured and transparent stages or steps in developing evidence-informed policy (Adams & Sandbrook, 2013; Norris et al., 2021). Whilst there is some diversity in these frameworks, it is clear that the following steps are key to a good process:

- Defining the problem (Christie et al., 2022; Salafsky et al., 2022)
- Gathering and assessing evidence (Bowen & Zwi, 2005; Christie et al., 2022; Salafsky et al., 2022)
- Integrating evidence into wider policy-making practices (Bowen & Zwi, 2005; Christie et al., 2022; Salafsky et al., 2022).

Methods for evidence synthesis

Frameworks can be used to understand the ‘problem’ and use evidence to develop policy and there is a range of methods and tools for gathering, assessing, and synthesising evidence. Evidence synthesis informs the user of what is known from research and comes in a variety of forms (OECD, 2020). Evidence synthesis is a set of methodological approaches for systematically identifying, screening, appraising the quality, and synthesising primary research evidence (Macura et al., 2019).

Choosing the appropriate method of evidence synthesis depends on the type of review question, purpose of the review, type of data, and availability of expertise, time and funding (Macura et al., 2019).The main methods of evidence synthesis are:

- Systematic reviews
- Rapid reviews
- Traditional reviews
- Umbrella reviews (Reviews of reviews)
- Systematic reviews (Scoping reviews)
- Meta-analyses

The diversity of approaches to evidence synthesis provides a wide range of options suited to different decision-making contexts. When choosing among multiple methods of evidence synthesis, time and resource constraints are an important consideration, but the scope and type of question being asked, and the level of certainty required from the synthesis, are equally important.

Simple tools are available to assist with the selection of the most appropriate evidence synthesis method to use given the circumstances (size of the team, time constraints, nature of the research question, and importance of the decision to be made; see for example Cook et al., 2017 and Sutherland et al., 2021). Cornell University Library has a useful flowchart which the Ministry could adapt as part of the process for assessing the strength of scientific evidence.

Key steps in an evidence synthesis

The key steps involved in a systematic review; the most robust form of evidence synthesis; are:

1. Planning the synthesis and developing the question
2. Developing a protocol
3. Conducting a systematic search
4. Conducting a systematic eligibility screening
5. Data coding and extraction
6. Critical appraisal of the eligible resources
7. Data synthesis
8. Interpreting findings and reporting.

For the most part these steps are also applicable to rapid reviews and systematic maps, with the key differences outlined in Table 1.

Frameworks for environmental evidence synthesis

The report describes a series of different frameworks incorporating guidelines and tools to deliver each of the eight key steps of the evidence synthesis process for systematic / rapid reviews or systematic maps in fields of ecology, environmental management and conservation:

- Collaboration for Environmental evidence framework
- Conservation Evidence framework
- Eco Evidence framework
- US Environmental Protection Agency framework

A comparison of the frameworks is provided in Table 4, with Table 5 assessing the extent to which each framework is consistent with the criteria outlined earlier in this summary.

Other tools for evidence synthesis

In addition to the frameworks described above, there is a range of additional tools which are not part of a framework but can be incorporated into an evidence synthesis. The report provides an overview of some of the tools developed for the fields of ecology, environmental science and conservation and of other tools which are widely used in other fields. These tools can be used either to appraise individual studies, or to appraise reviews.

Communicating the findings of syntheses to policy makers

Environmental policy decisions in New Zealand can range from those made at the central government level (e.g., developing or amending a national policy statement) to plan making at a regional or district council level. This may mean that there will be constraints on the time and resources available to synthesise evidence to inform a decision. Policy makers generally work in fast paced environments (and under fast-paced processes), are often time-poor, and can be bombarded with information from a range of sources (Wood, n.d.). Collectively, these factors combine to produce a challenging context within which the communication of the findings of evidence synthesis needs to be made to inform a decision.

The way in which these findings are presented can vary from relatively dense technical papers through to one- or two-page non-technical summaries. Communicating findings in a way that enables policy makers to get a quick overview of the review while also providing links to additional information is likely to be particularly valuable.

Providing information about the findings across the range of interventions assessed in a consistent manner will also assist decision makers to more easily compare and contrast the levels of evidence supporting a range of interventions to achieve a particular outcome.

CONTENTS

Appendices

Figures

1. Introduction

1.1 Context

It is widely accepted that scientific evidence should inform the development of effective policy. Evidence is important for both understanding complex interactions, and for politically justifying policy and management decisions (Kadykalo et al., 2021). However, the process of developing evidence-based policy is a challenge experienced in all fields of policy making. This is particularly the case in environmental policy due to the political prominence and broad sectoral reach of policies, tensions raising from policy making at different levels, the wide range of spatial and temporal scales, the complex, uncertain and contested nature of the evidence base, and the irreversibility of damages (Macura et al., 2019; Reed & Meagher, 2019). Uncertainty, gaps and biases in understanding evidence can impact decision making and increase the risk of poor outcomes. Both the quality of evidence and / or the uptake of evidence can influence the quality of policy, guidelines and decision making.

For example, mixed and contrasting evidence (poor-quality evidence), often based on common sense approaches, hindered the formulation of clear guidelines to prevent sudden infant death syndrome (SIDS) for most of the 21st century (Gilbert et al., 2005). The safest sleeping position for infants (on the back) was not recommended consistently until 1995. This was despite the fact that sufficient information in support of this recommendation had been available since the early 1970s, as found by a systematic review published in 2005 (Gilbert et al., 2005). Had systematic reviews been common practice in the 1970s, there could have been an earlier recognition of the risks of sleeping on the front, which could have prevented over 10,000 infant deaths in the UK and at least 50,000 in Europe, the USA, and Australasia (Gilbert et al., 2005).

In the field of environmental management and conservation, many ecological mitigation measures to avoid / minimise biodiversity losses in the face of new infrastructure developments have been found to lack a solid evidence base (Hill & Arnold, 2012; Hunter et al., 2021; Singh et al., 2020). For example, a review of 65 ecological mitigation measures recommended to address the impacts of housing developments in the UK found that only 56% of these measures were supported by citing published guidance. In addition, a further review of the published guidance found that less than 10% of the evidence cited by the guidance documents was derived from empirical evaluations of the effectiveness of the measures that were recommended (Hunter et al., 2021).

More generally, most recently-published evidence syntheses in the field of environmental management and conservation have been found to be of low reliability to inform decision making because important information describing methodology and results are frequently missing (O'Leary et al., 2016; Pullin et al., 2022). This consistent lack of transparency and methodological rigour is particularly disappointing since rigorous review methodology and reporting standards are available (see Section 2.2). Lack of adherence to the available standards could be due to lack of time or funding, lack of methodological awareness, disagreement over the need for some criteria, or meeting high standards of conduct being regarded as disproportionate to the impact of the evidence synthesis being undertaken (Pullin et al., 2022).

In New Zealand, the National Policy Statement for Freshwater Management (NPS-FM) is an example of environmental policy that has been highly scrutinised. The NPS-FM has been

amended multiple times (in 2011, 2014 and 2017, and most recently in 2020) and is currently under review. The process for developing the NPS-FM 2020 has been analysed by (Koolen-Bourke & Peart, 2022) who attribute the strength of the latest policy from its predecessors as the inclusion of more diversity on the advisory group, the exclusion of economic considerations from the consideration of science, and the provision of independent reports setting out the science advice.

Even when good-quality evidence is available, myriad factors can limit its uptake by managers, decision-makers and other end-users (Walsh et al., 2019). For example, many of the measures included in the Common Agricultural Policy (CAP) of the European Union to reduce the environmental impacts of agriculture were known to be ineffective even before being implemented, while more suitable options were discarded because the available evidence was not taken into account (Dicks et al., 2014; Pe'er et al., 2020; Sutherland, 2022).

There are countless examples of poor conservation outcomes due to the failure to account for the evidence on constraints and limitations to the effectiveness of conservation actions, and surveys of conservation practitioners have revealed a non-systematic use of scientific evidence, primarily because of the high volume and dispersed nature of the information (Pullin & Knight, 2005; Walsh et al., 2015).

1.2 Challenges associated with implementing evidence-informed environmental policy decisions

Current barriers that may constrain the effective use of evidence in policy development include accessibility, relevance and applicability, organisational capacity, resources and finances, time constraints, and poor communication and dissemination skills between scientists and decision-makers (Cooke et al., 2023; S. J. Nichols et al., 2017). Some of the other key challenges that set the context for this issue are outlined below.

Trust in political and scientific authorities

Public trust in political and scientific authorities has declined significantly over the past several decades across most democratic societies (Hendriks et al., 2016). Causes are complex, including resurgent populism, political polarisation, culture war struggles, and the expansion of anti-establishment alternatives and social media (Hosking, 2019). Pushing back against these trends to build public trust and acceptance of evidence and evidence syntheses will be challenging (Cooke et al., 2023) but is part of the context within which this study has been undertaken.

In addition to this general context, public controversy or mistrust are more likely when the stakes are higher and decisions may lead to losses of income or opportunity (potentially in the short-term), as is often the case with environmental governance (Sarewitz 2004 in Cooke et al. (2023)).

Dealing with vested interests

Politicians, public officials, experts, stakeholders, and the general public may all have varying degrees of interest in the process of policy making, depending on the nature of the policy problem. Sharman and Holmes (2010, in Kano & Hayashi (2021)) observed that the political aspects of the production and use of evidence are sensitive issues, and that policymakers need evidence that is more useful in a political context. Simultaneously, policymakers' motivations (which might be driven by the election cycle, for example) have the inextricable potential to lead to the arbitrary cherry-picking of evidence. This results in "policy-based evidence gathering",

which contradicts evidence-based policy making. Policymakers will likely face the weighing of the opinions of scientists and their evidence over democratic values when making decisions (Laing and Wallis, 2016; Cvitanovic et al., 2015; Garvin 2001 in Kano & Hayashi, 2021).

Complexity of environmental problems

Environmental problems are often cited as being highly complex in the literature (Haug et al., 2010; Macura et al., 2019) and present particular challenges for evidence-informed policy and practice. This can be due to:

- the political prominence and broad sectoral reach of policies;
- tensions raising from policy making at different levels;
- the wide range of spatial and temporal scales;
- the complex, uncertain and contested nature of the evidence base; and
- the irreversibility of damages (Kano & Hayashi, 2021; Macura et al., 2019; Reed & Meagher, 2019).

Nichols et al. (2017) describe the need for sound decision-making on environmental matters to be informed by an understanding of cause-effect relationships but notes how this is challenging.

Types of evidence

One of the challenges with understanding evidence-informed policy is defining the types of information that can be used to evidence a policy decision (Adams & Sandbrook, 2013; Christie et al., 2022; Cooke et al., 2023; Salafsky et al., 2019). While there are many ways to describe evidence, it is usually one of three types: research evidence, contextual evidence and experiential evidence (Figure 1, from Superu (2018)).

Research evidence, sometimes referred to as ‘scientific evidence’ is a body of information based on numerical (quantitative) or descriptive (qualitative) facts and can include systematic reviews, meta-analyses, randomised control trials, and surveys.

Contextual evidence is from process or formative evaluation, surveys or census, and longitudinal / cohort studies, while experiential evidence stems from case studies and focus groups, oral histories and interviews, and user feedback. Strong policy brings together all three types of evidence to form high-quality and relevant advice (Superu, 2018).

Figure 1: The inter-relationship between the three types of evidence (Superu, 2018).

1.3 The policy cycle in New Zealand

Warner (2022) describes the policy cycle in New Zealand as being structured around five key stages as illustrated in Figure 2:

1) Agenda setting
2) Policy formulation
3) Decision-making
4) Implementation
5) Evaluation

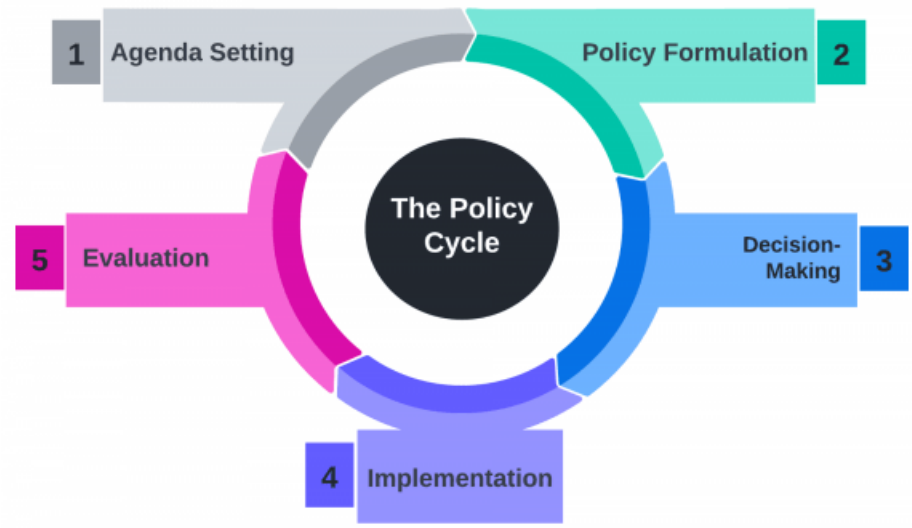


Figure 2: Policy cycle (Source: Warner (2022)).

Within these five stages of the policy cycle there may be additional steps. For example, problem identification, information gathering and research may form part of stage 1; inter-agency consultation and tendering advice with recommendations to government may occur prior to stage 3; political consultation and negotiation, and legislation formation can be important steps occurring prior to stage 4 (Office of the Ombudsman, 2019).

The policy cycle is theoretical and assumes that policy-making is a linear process. In reality, the order of these five stages above may vary and steps may be repeated (Office of the Ombudsman, 2019). The policy process can be defined as complex, multifactorial and nonlinear (INASP, 2016).

In practice, the policy cycle is applied across various levels of government in New Zealand, including central, regional and local government. Public consultation and engagement may be undertaken in some form, and at any and all stages of the policy cycle.

The 'scale' of policy-making processes can vary significantly, for example from the development of a new piece of national direction under the Resource Management Act to the review of a regional land and water plan, or the ongoing monitoring of the state of the environment for the purposes of fulfilling the requirements under the Environmental Reporting Act 2015.

In many instances, associated processes may be described in legislation and may include statutory timeframes, which can place limitations on the time available to obtain, review, and synthesise evidence. The level of resources available to the various agencies who support or lead the delivery of these processes can vary and may often result in constraints on the 'ideal' approach that might otherwise be followed to inform an evidence-based policy decision.

Within this diverse policy-making landscape, the challenge of ensuring that scientific evidence appropriately informs policy decisions becomes equally complex.

In light of the context and challenges outlined above, the Ministry for the Environment (the Ministry) has identified a need to ensure scientific evidence is assessed appropriately and informs the development of effective policy, to avoid the misuse of evidence, which can result in poor policy-making and negative consequences for environmental management.

The objective of this project is to make recommendations that will inform the potential development of a repeatable, transparent process for assessing the strength of scientific evidence for the development of 'science-informed policy' that relates to environmental matters. The Ministry engaged Boffa Miskell to undertake this project.

This remainder of this report sets out an overview of the approach we have taken to deliver this project, and a presentation of findings. The report is concluded with a series of recommendations that the Ministry may use to develop a process for assessing the strength of scientific evidence in ongoing environmental policy-making.

1.4 Literature review methods

Boffa Miskell was commissioned by the Ministry to undertake a thorough and critical literature review of approaches to assess the strength of scientific evidence for informing environmental policy development. The scope of the literature review included identifying and evaluating the key principles, frameworks and methods used within New Zealand and internationally.

As the assessment of evidence in environmental policy is, internationally and within New Zealand, at an earlier stage of maturation compared to its application in other fields (e.g., education and health), there is also reference made to frameworks and learnings from fields outside of environmental science.

A range of key search terms derived from the project objective were used to search databases (Google, Google Scholar) to gather initial literature material. Key search terms included variations of the following:

> *Approaches to assessing the strength of scientific evidence / assessing evidence for policy / assessing science evidence / evidence-informed policy / evidence-based policy / evidence for environmental policy / evidence frameworks / evidence thresholds.*

Reference lists from literature found were used to identify other relevant literature using similar search terms. A targeted review of overseas government websites was also carried out to obtain applied literature, such as guidelines for evidence-informed policy. Key themes were extracted, categorised, and analysed, then discussed among the Boffa Miskell researchers, reviewers, and a Challenge Group, to shape the literature review through an iterative process.

1.4.1 Assessing the strengths and weaknesses of the frameworks and methods

The findings of the literature review were used to describe various approaches used for the appraisal and synthesis of scientific evidence and to assess the strengths and weaknesses of each approach. Strengths and weaknesses of each approach were assessed in relation to the following criteria established by the Ministry to identify an approach that:

- Allows for an 'absolute' assessment of evidence.
- Is transferable across policy questions in different environmental domains and at different levels of government.
- Is repeatable and allows for follow up assessments in the future.
- Can be applied to different stages of the policy cycle.
- Ensures a transparent evaluation of scientific evidence and reduces the likelihood of selection biases.

In considering the relative strengths and weaknesses of the various approaches in the literature, we have also been cognisant of some of the key challenges that can limit evidence-informed policy, such as those described by Cooke et al. (2023) and Nichols et al. (2017):

- accessibility
- relevance and applicability
- organisational capacity
- resources and finances
- time constraints
- poor communication and dissemination skills between scientists and decision makers.

This literature review focuses on scientific research evidence and does not include the assessment of other forms of evidence such as mātauranga Māori as it would be more appropriate to address this in a separate piece of work.

2. Findings

2.1 Frameworks used for evidence-informed policy

Frameworks are the predominant tool for ensuring a principled process of evidencing a policy concern by guiding the user through steps / stages to ensure a transparent, credible and legitimate process (Schwartz et al., 2018). Some refer to these frameworks as evidence-to-decision frameworks (Norris et al., 2021) or Decision Support Frameworks (HM Treasury, 2020; Schwartz et al., 2018).

The literature consistently describes how introducing a framework guides the user through a set of structured and transparent stages or steps in developing evidence-informed policy (Adams &

Sandbrook, 2013; Norris et al., 2021). Whilst there is some diversity in these frameworks, it is clear that the following steps are key to a good process:

- defining the problem (Christie et al., 2022; Salafsky et al., 2022)
- gathering and assessing evidence (Bowen & Zwi, 2005; Christie et al., 2022; Salafsky et al., 2022)
- integrating evidence into wider policy-making practices (Bowen & Zwi, 2005; Christie et al., 2022; Salafsky et al., 2022).

The Evidence-to-Decision Tool framework (Figure 3) is a good example of this, where the first step is to define the problem, and the following steps involve gathering the evidence and using it to make a decision. Within these stages there are key nuances. For instance, Salafsky et al. (2022) say that a distinct step to determine the level of confidence in the evidence is needed (i.e., evidence threshold). Pullin & Knight (2003) describe a two-step process of producing systematic reviews and then making this available to the decision maker. These processes assume that evidence-gathering and analysis are undertaken independently of decision making.

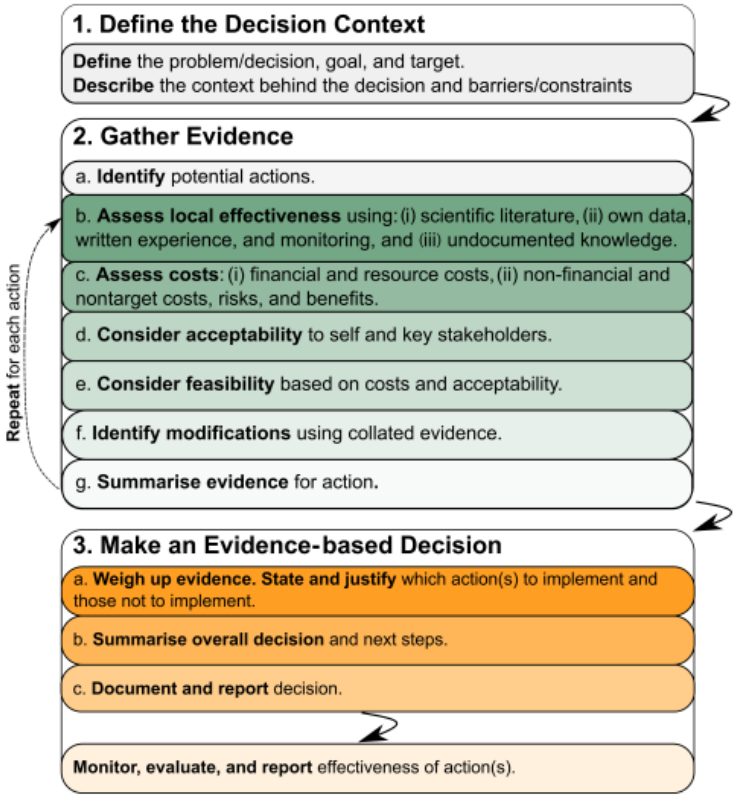


Figure 3: The evidence-to-decision tool is an example of a framework used for evidence-informed policy (Source: Christie et al., 2022).

The application of frameworks sees the use of methods such as strategic foresight, systematic conservation planning, structured decision making, open standards and evidence-based practice (Christie et al., 2022; HM Treasury, 2020; Schwartz et al., 2018).

The strengths of frameworks have been documented as enabling users to make the rationale and process behind making decisions explicit, documentable and transparent (Christie et al.,

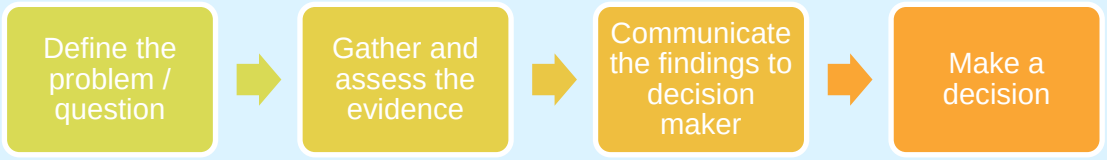
2023a; Norris et al., 2021), by presenting a formalised approach to combing evidence from diverse sources, whilst being versatile in their application (Christie et al., 2022; Schwartz et al., 2018).

Nevertheless, critiques of frameworks are that they are only a guide to users (i.e., encourages but does not enforce) to document, report or share the decision-making process. Most importantly, frameworks do not stop decision making bias (Christie et al., 2022). Norris et al. 2021) warn that adopting frameworks from other disciplines (such as from clinical medicine) is challenging and there are important differences in the assessment of quality evidence for environmental policy versus health sciences, and in translating to policy recommendations. The reality of environmental policy-making means that linear frameworks used in clinical medicine are not as easily applied in practice; environmental policy-making is much more complex and messy than health policy-making (Adams & Sandbrook, 2013).

RECOMMENDATION *1*. FRAMEWORKS USED FOR EVIDENCE-INFORMED POLICY

The use of a framework to guide researchers and policymakers through a set of structured and transparent stages or steps in developing evidence-informed policy is recommended, and that the framework should include the following steps as illustrated below:

1) Define the policy problem / question
2) Gather and assess the evidence (evidence synthesis)
3) Communicate the findings of the evidence synthesis to the decision maker
4) Make a decision



This recommendation supports:

- ☑ A transferable process
- ☑ A repeatable process
- ☑ A transparent evaluation of scientific evidence, aiming to reduce bias
- ☑ A high-level framework within which the steps undertaken can be tailored to the time, capacity and resources available.

2.2 Methods for evidence synthesis

Frameworks can be used to understand the ‘problem’ and use evidence to develop policy and there is a range of methods and tools for gathering, assessing, and synthesising evidence. Evidence synthesis informs the user of what is known from research and comes in a variety of forms (OECD, 2020). Evidence synthesis is a set of methodological approaches for systematically identifying, screening, appraising the quality, and synthesising primary research evidence (Macura et al., 2019).

Over the last 20 years, there has been a proliferation of methods for evidence synthesis driven by the need to ensure the product is fit-for-purpose within the decision-making context. This growth has often not been well coordinated within and between disciplines, leading to confusion among both scientists and practitioners about the strengths and weakness of different approaches and the circumstances in which they are likely to be most appropriate (Cook et al., 2017).

Choosing the appropriate method of evidence synthesis depends on the type of review question, purpose of the review, type of data, and availability of expertise, time and funding (Macura et al., 2019). Some of the most common and more widely used evidence synthesis methods are outlined in Section 2.2.1. Further information on selecting the most appropriate method is provided in Section 2.2.2.

2.2.1 Overview of the main methods

Systematic reviews

Systematic reviews employ a comprehensive search approach as a defining feature. This provides a comprehensive assessment of evidence within a field and a rigorous method to assessing studies and referencing (Sutton et al., 2019; Collaboration for Environmental Evidence, 2022). Systematic reviews are considered the most robust method for reviewing, synthesising and mapping existing evidence on a particular topic, and for that reason are resource and time intensive (OECD, 2020). In addition, systematic reviews rely on a substantial body of evidence and do not work as well when there is limited evidence available (HM Treasury, 2020).

Key features of systematic reviews are transparency and reproducibility. Formal guidance and standards for systematic reviews are well established and are provided by organisations such as Cochrane Collaboration (in the field of healthcare interventions)[1], the Campbell Collaboration (in the fields of education, social welfare, and crime and justice)[2], and the Collaboration for Environmental Evidence (in the field of environmental management and conservation)[3].

Rapid reviews

Rapid reviews (also known as rapid evidence assessment and rapid evidence synthesis) are quick reviews of the evidence when resources are limited, or the topic is urgent. Rapid evidence assessments may take a variety of forms, but typically follow the same processes as systematic reviews, with stages omitted or abbreviated. While some types of rapid review may abbreviate the search process, for others the time savings are made elsewhere in the process, for example through the removal or simplification of the appraisal, synthesis, or analysis stages. In essence, rapid reviews offer a flexible template, but any deviation from the conventional systematic review methods should be well documented (Sutherland, 2022; Sutton et al., 2019).

[1] https://www.cochrane.org/

[2] https://www.campbellcollaboration.org/

[3] https://environmentalevidence.org/

Traditional reviews

Traditional reviews include critical reviews, integrative reviews, narrative reviews, narrative summaries, and state of the art reviews. Traditional reviews employ bibliographic database searching, but they are not as explicit in their methods as systematic reviews, although there is a move to be more systematic as transparent reporting is increasingly expected (Haddaway et al., 2015; Sutton et al., 2019). Traditional reviews can provide an alternative to systematic / rapid reviews when time and resources are limited, and when the systematic approach is unsuitable or unnecessary (Haddaway et al., 2015).

Umbrella Reviews (Reviews of Reviews)

Umbrella reviews (also known as reviews of reviews) bring together multiple reviews to map and synthesise an existing evidence base. Reviews of reviews follow the same methodological and reporting standards as systematic reviews, but they focus on the findings of systematic reviews or other evidence syntheses rather than on individual studies (Sutton et al., 2019).

In some situations, there may already be existing evidence syntheses available. For example, Collaboration for Environmental Evidence (CEE) provides an open-access database of Evidence Reviews (CEEDER[4]) that collates evidence syntheses relevant to environmental management, policy interventions and anthropogenic impacts on the environment. The dataset aims to help policy makers, managers, funders, and the public find reliable evidence to inform their decision making in environmental management.

Other sources of evidence syntheses in the field of environmental management and conservation include Conservation Evidence, which provides synopses of reviews of the effectiveness of conservation actions[5], and the Nature-based Solutions Initiative, which provides a systematic map of evidence on the effectiveness of nature-based interventions for climate change adaptation[6].

Systematic maps (Scoping reviews)

Systematic maps (also known as mapping reviews and evidence maps) are particularly useful to understand the extent and nature of the evidence base on a broad topic area. They differ from systematic reviews in that they do not describe the findings of a certain body of evidence. Systematic maps can help describe the distribution of existing evidence, highlighting areas of significant research effort and where key gaps exist. This can provide information to guide research, prioritise evaluation, and illustrate where there may be inadequate information to inform decision making (Sutherland, 2022; Sutherland & Wordley, 2018; Sutton et al., 2019).

Meta-analyses

Meta-analyses comprise a set of statistical methods for combining the magnitude of outcomes (effect sizes) across different studies addressing the same research question. Meta-analyses pool data form multiple studies and analyse them together to assess the overall magnitude and consistency of a given effect (Borenstein, 2009; Gates, 2002; Gurevitch & Hedges, 2020; Koricheva, Gurevitch, et al., 2013; Nakagawa et al., 2023).

Meta-analyses are key for the data synthesis stage of systematic review (see Sections 2.2.2, 2.2.3, and Appendix 1) and are also widely used as part of studies not involving a systematic

[4] https://environmentalevidence.org/ceeder-search/

[5] https://www.conservationevidence.com/synopsis/index

[6] https://www.naturebasedsolutionsevidence.info/

search of all evidence. While meta-analyses are powerful tools to bring together the evidence from multiple studies, there is increasing awareness of large changes in the results of meta-analyses as the evidence on a given topic keeps accumulating (i.e., temporal instability in magnitude and significance of the reported effects; (Brisco et al., 2023; Koricheva, et al., 2013).

2.2.2 Selecting the best method for the circumstances

The diversity of approaches to evidence synthesis provides a wide range of options suited to different decision-making contexts. When choosing among multiple methods of evidence synthesis, time and resource constraints are an important consideration, but the scope and type of question being asked, and the level of certainty required from the synthesis, are equally important.

The acceptable level of certainty for decision-making is highly context dependent. For example, some decisions are irreversible (or have greater consequence) and require a greater level of certainty. Greater levels of certainty can be provided by more comprehensive and systematic approaches, such as systematic reviews. However, for decisions with lower levels of consequences, a higher level of uncertainty may be acceptable and methods with less stringent requirements (and, therefore less time and cost) than systematic reviews may be considered (Cook et al., 2017; Salafsky et al., 2019; Sutherland et al., 2021).

Despite the abundance of synthesis methods (with inconsistent nomenclature contributing to make the choice overwhelming), systematic reviews, rapid reviews and systematic maps are the tools that appear best developed and more widely used.

Systematic maps are often preliminary syntheses of the evidence relating to a broader question, while systematic reviews (or their abbreviated rapid form) aim to answer a specific question by collating and synthesising findings of individual studies in order to produce an aggregate measure of effect or impact (Sutton et al., 2019; Collaboration for Environmental Evidence, 2022).

A **systematic review** may be appropriate to:

- Measure the effectiveness of an intervention on a specific population / natural system
- Measure the impact of an activity on a specific population / natural system
- Assess the quantity and quality of research that has been conducted on a specific topic.

A **systematic map** may be appropriate to:

- Describe the evidence base for a given topic
- Answer broad questions such as *“what interventions have been used to decrease the impact of commercial fishing on marine biodiversity?”* or *“what are the impacts of agri-environment schemes on farmland biodiversity?”*

Standards and protocols for systematic reviews (or their abbreviated rapid form) and systematic maps are discussed in Section 2.2. Standards and protocols ensure that use of these methods is rigorous, objective and transparent to minimise bias and work toward consensus among stakeholders on the status of the evidence base. Having a transparent and well-defined framework is essential in situations when there are opposing views about the effectiveness of interventions or impact of actions, or when the effectiveness of interventions needs to be weighed against their costs.

However, it is also important to consider that systematic / rapid reviews and systematic maps may not always be appropriate or required, for example when:

- The question is poorly defined or too complex
- The question is too simple (e.g., has a certain species been recorded in a certain area)
- The question is of lo stakeholder and scientific interest and can be satisfactorily answered with methods that are less rigorous and less costly
- There is insufficient good quality evidence available.

Simple tools are available to assist with the selection of the most appropriate evidence synthesis method to use given the circumstances (size of the team, time constraints, nature of the research question, and importance of the decision to be made; for example Cook et al., 2017 and Sutherland et al., 2021). Cornell University Library has a useful flowchart, reproduced in Figure 4 below, which the Ministry could adapt as part of the process for assessing the strength of scientific evidence.

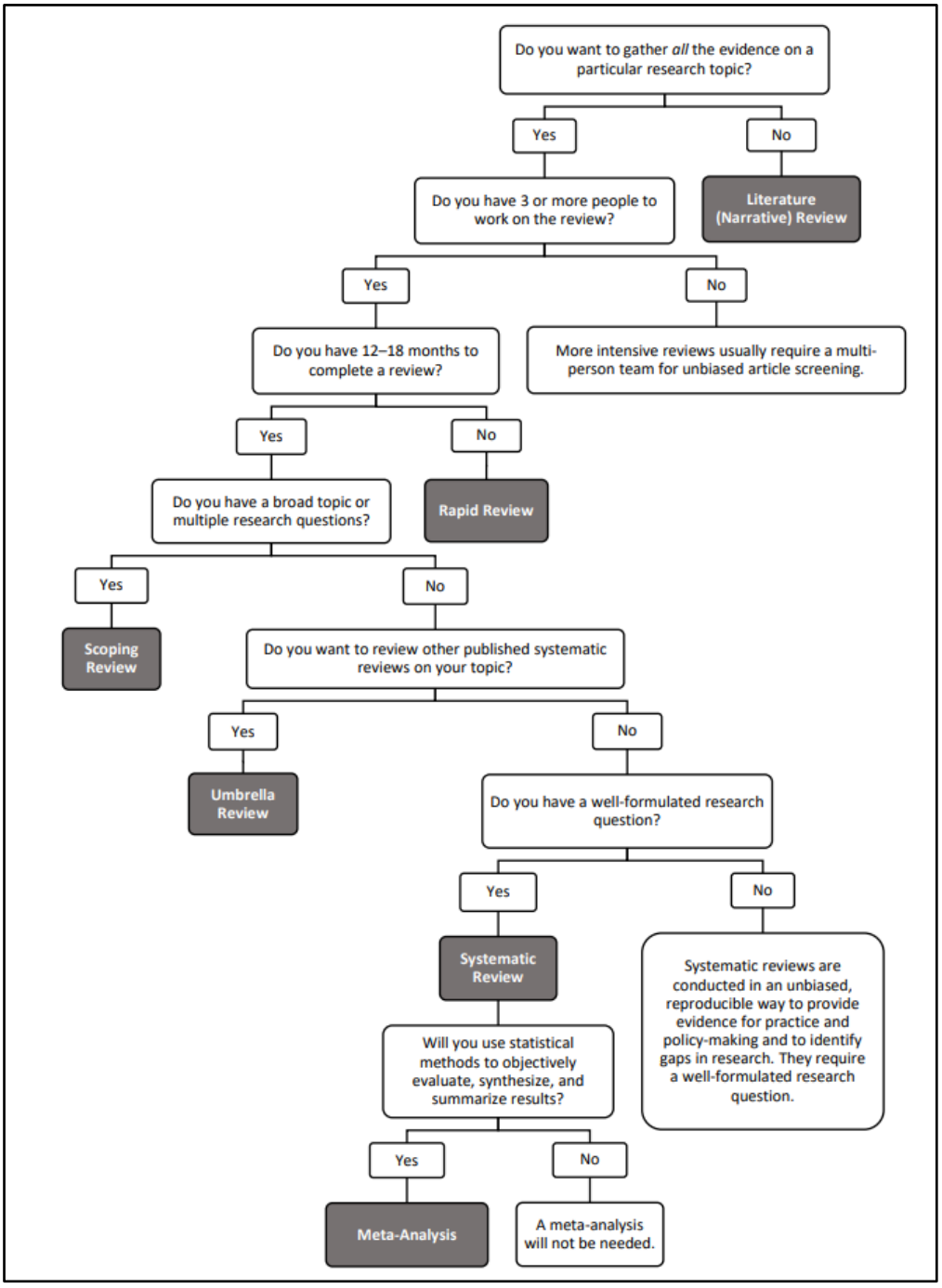


Figure 4: Flow chart to assist in selecting the most appropriate evidence synthesis tool (Cornell University Library[7]). Note that the Scoping Review is equivalent to Systematic Maps.

[7] https://guides.library.cornell.edu/ld.php?content_id=52561085

RECOMMENDATION 2. SELECTING THE EVIDENCE SYNTHESIS METHOD

It is recommended that the following factors are used to inform a decision about the most appropriate method of evidence synthesis to use:

1) The nature of the question or problem
2) The level of certainty required from the synthesis
3) The time and resources available

Simple flow charts (see Figure 4) can help with the choice among different evidence synthesis methods.

This recommendation supports:

- ☑ Using a repeatable process to decide which evidence synthesis method to use
- ☑ Selecting the best method to apply at different stages of the policy cycle (in response to the time, resources and expertise available at each stage)

2.2.3 Key steps in evidence synthesis

The key steps involved in a systematic review; the most robust form of evidence synthesis; are:

1. Planning the synthesis and developing the question
2. Developing a protocol
3. Conducting a systematic search
4. Conducting a systematic eligibility screening
5. Data coding and extraction
6. Critical appraisal of the eligible resources
7. Data synthesis
8. Interpreting findings and reporting.

Each of these steps is described below (and in further detail in Appendix 1) based on the guidance provided by Collaboration for Environmental Evidence (CEE) for systematic reviews (Collaboration for Environmental Evidence, 2022)[8]. For the most part, these steps are applicable also to rapid reviews and systematic maps (the evidence synthesis tools better developed and more widely used along with systematic reviews) with some key differences in relation to each of the steps which are outlined in Table 1 and discussed in the following sections.

Importantly, some of the methods and principles outlined below are also applicable to traditional reviews. In particular, careful documentation of all methods and limitations to ensure procedural objectivity, consistency and transparency would be a valuable and relatively straightforward way to improve the quality and reliability of traditional reviews (Haddaway et al., 2015; Sutton et al., 2019).

[8] https://environmentalevidence.org/information-for-authors/

Table 1: Key evidence synthesis steps in systematic reviews, rapid reviews and systematic maps.

Evidence synthesis process	Systematic reviews	Rapid reviews	Systematic maps
1. Planning and question	Requires a specific, detailed question	Requires a specific, detailed questions	Suitable for broader questions
2. Protocol	Mandatory	Mandatory	Mandatory
3. Search	Systematic search	Systematic search shortened with pre-determined restrictions	Systematic search
4. Screening	Systematic screening	Systematic screening shortened with pre-determined restrictions	Systematic screening
5. Data coding and extraction	Metadata coded and outcomes measures (effect size) extracted	Metadata coded (with pre-determined restrictions to shorten the process) and outcomes measures (effect size) extracted	Metadata coded
6. Appraisal	Mandatory	Mandatory with pre-determined restrictions to shorten the process	Optional
7. Data synthesis	Narrative and quantitative synthesis (meta-analysis)	Narrative and quantitative synthesis (meta-analysis)	Narrative synthesis and data visualization
8. Reporting	Narrative and (where possible) quantitative answer to the question	Narrative and (where possible) quantitative answer to the question	Narrative description of the evidence base

Planning a synthesis and developing the question

The primary aim of planning a synthesis and developing the question when undertaking an evidence synthesis is to inform the goals and structure of the review and to ensure that the process of evidence synthesis is as free from bias as possible. At this time, the involvement of relevant stakeholders in steps a) to e) described below is essential to inform the goals and structure of the review and to ensure that the process of evidence synthesis is as free from bias as possible. For complex reviews, subject experts or advisory panels may be consulted during the planning stage.

a) Defining the question to be answered.

   i. Key elements for making a question suitable for evidence suitable for evidence synthesis can be referred to using the Population, Intervention, Comparator and Outcomes (PICO) or Population, Exposure, Comparator and Outcomes (PECO) acronyms. Further information on these systems and how to use them is set out in Appendix 1.

   ii. Questions that are specific, well defined, and relatively simple are well suited to be addressed via systematic reviews.

iii. Questions that are more open ended / broad in nature are better suited to systematic maps.

b) Undertake a preliminary scoping of the evidence to guide the selection of the evidence synthesis method and the development of the review protocol.

c) Estimate resourcing (budget and personnel) requirements and timelines.

d) Select the evidence synthesis method that will be used (systematic review, rapid review, or systematic map). Further details on making this selection are set out in Section 2.2.2.

e) Assemble the review team based on the following principles:

iv. Systematic reviews usually require a multidisciplinary team, led by a Lead Reviewer.

v. The team may include subject matter experts working alongside review and synthesis methodology experts.

vi. Conflicts of interest should be avoided, but declared if they arise.

Developing a protocol

The protocol for the evidence synthesis is an independent document to be prepared before the synthesis is conducted. The protocol serves as a guide and reference to the conduct of the synthesis, which should reflect the views of all the parties involved in the planning phase (i.e., the commissioner of the work, the ultimate users of the evidence, the stakeholders in the process and the review team).

The protocol is essential to minimise reviewer bias. Any diversion from the protocol during the synthesis process is discouraged. However, when changes to the original methodology are necessary, these must be compulsorily recorded and motivated. This is particularly important to maintain transparency and repeatability, as well as the confidence of users of the evidence and stakeholders. Further details on the purpose of the protocol and what it should address are set out in Appendix 1.

Conducting a search

Searches should be transparent and reproducible. In practice, it is unlikely that absolutely all the relevant literature can be identified during the search, but a key requirement is to try to gather as much of the available evidence as possible to minimise bias in the findings. Any limitations of the search, such as lack of access to or inability to use some literature (for example, because of a language barrier) should be clearly reported.

Enlisting an information specialist in the review team is recommended to establish an efficient search strategy. A good search strategy can also make a substantial difference to the time and cost of a synthesis. In addition, because of the systematic aspect of the searching and the need to keep careful track of the findings, review teams should, when possible, include librarians or information specialists.

The search should ensure the following key principles are adhered to (with further details set out in Appendix 1):

a) Establishing a test-list

b) Search errors are avoided

c) Search biases are avoided (however some limitations might be placed when undertaking a rapid review – i.e. limiting the search to certain languages, time periods, or geographical locations)

d) Identifying search terms and developing search strings

e) Searching different types of sources (the range of sources may be narrowed if a rapid review is being undertaken)

f) Having defined criteria for when to stop the search

g) Keeping track of the search details

h) Reporting all relevant information about the search in the final evidence synthesis report

i) Updating or amending the search to reflect new evidence as it becomes available.

For rapid reviews, measures to speed up the search should be documented and justified, and may include:

- Adopting date, language, geographical limitations
- Searching only key databases.

Screening the evidence to establish whether the resources identified are relevant to the question

Eligibility criteria are used as part of a systematic screening process to establish whether the resources identified by the search are relevant for answering the question driving the evidence synthesis. Both the eligibility criteria and the screening process should be planned in advance and must be specified in the evidence synthesis protocol. Further information on this step is set out in Appendix 1.

For rapid reviews, measures to speed up the screening should be documented and justified, and may include:

- Using only one screener, but as many references as possible should be dual-screened and the consistency of screening decisions should be tested.
- Using eligibility criteria that place emphasis on higher validity study designs.

Data coding and extraction

Data coding and extraction refer to the process of systematically extracting relevant information from the resources retained following the screening process.

- Data coding is the recording of relevant characteristics (meta-data) of the study such as when and where the study was conducted and by whom, as well as aspects of the study design and conduct.
- Data extraction is only required for systematic reviews and refers to the recording of the results of the study (e.g., effect size, means and variances or other important findings).

Further details are available in Appendix 1. For rapid reviews, measures to speed up data coding and extraction should be documented and justified, and may include:

- Using only one reviewer, but a sample of resources should be examined independently by two reviewers to test for consistency.
- Limiting coding and extraction to data necessary for the synthesis.

Critical appraisal of the eligible resources

In the critical appraisal stage, the resources retained following the screening process are assessed for their reliability for answering the question motivating the evidence synthesis. As the quality of scientific evidence varies considerably, the critical appraisal step is essential to identify the flaws and limitations of the evidence being used so that these can be considered when drawing the conclusions of the synthesis.

There are two key elements that need to be considered when appraising the evidence:

- **Internal validity**. Internal validity refers to the extent of bias in the results of an individual study due to flaws in study design or conduct. The extent of bias can be inferred by examining the study design and methods to determine whether adequate steps were taken to protect against bias.
- **External validity**. Whilst internal validity is a specific property of an individual research study, external validity is context dependent. External validity is the extent to which the results of an individual study can be generalised and applied to other circumstances. This includes the suitability of the findings of a study for answering the question being addressed by the review. For example, how well do the results of control laboratory trials apply to answering a question related to effects / impacts occurring in the real world?

The critical appraisal process should be planned, and tested, while developing the protocol for the evidence synthesis. Key aspects of the appraisal process are outlined further in Appendix 1.

For rapid reviews, measures to speed up the appraisal of the evidence should be documented and justified, and may include:

- Using a risk of bias tool.
- Limiting risk of bias ratings only to certain form of bias depending on the outcomes of interest for end users and stakeholders.
- Using only one reviewer, but a sample of resources should be examined independently by two reviewers to test for consistency.

Data synthesis

Data synthesis refers to the collation of all relevant evidence identified in the review to answer the review question. A review should always have a narrative synthesis of the data, including a tabulation of key characteristics and outcomes of all the resources examined. For Systematic Reviews, if sufficient data is available in a suitable format, a quantitative synthesis, in the form of a meta-analysis, may also be planned.

As for all stages of a review, the data synthesis process should follow methods pre-specified in the review protocol, it should involve peer-review within the review team and accurately described in the final synthesis report.

Further details on the ways that data synthesis can be undertaken are set out in Appendix 1.

Interpreting findings and reporting

Evidence synthesis collates and synthesises data to present reliable evidence in relation to the review question. Authors should simply present the evidence to inform rather than offer advice. When reviews are inconclusive because there is insufficient evidence, it is important not to confuse “no evidence of an effect” (which may indicate the need for further research to build better evidence) with “evidence of no effect” (which instead would suggest that there is enough good-quality evidence to draw this conclusion).

RECOMMENDATION 3. KEY STEPS IN THE EVIDENCE SYNTHESIS PROCESS

1) The eight-step process outlined above is essential for systematic evidence synthesis and should always be followed.
2) Preparing and following a detailed protocol throughout all the steps is mandatory and it is what sets solid, systematic syntheses aside from other methods.
3) Subjectivity and technical constraints cannot be entirely removed from the evidence synthesis process. All the limitations of the study and the key decisions made by the review team (for example for the screening and appraisal of the evidence) should be carefully documented and explained in the protocol and in the final evidence synthesis report.
4) Any measure implemented to reduce time requirements as part of a rapid review should be thoroughly documented and motivated.
5) Any deviation from the original protocol should be thoroughly documented and motivated.
6) The first two steps of the process (covering general planning, formulation of the question and preparation of the protocol) should include consultation with relevant stakeholders (and with subject expert / advisory panels if needed) to inform the goals and structure of the review and to ensure that the process of evidence synthesis is as free from bias as possible.
7) Consider the possibility of adopting some of the methods and principles of systematic reviews when conducting a traditional review.

This recommendation supports:

- ☑ Transferability of a process across policy questions in different environmental domains and at different levels of government – there remains flexibility as to *how* each of the steps is delivered to adapt to the different circumstances.
- ☑ Implementation of a repeatable process that allows for follow up assessments in the future.
- ☑ Applicability to different stages of the policy cycle
- ☑ A transparent evaluation of scientific evidence – with the extent to which the likelihood of selection biases is reduced dependent on *how* certain steps are undertaken.

2.3 Frameworks for environmental evidence synthesis

The following sections of the report describe a series of different frameworks incorporating guidelines and tools to deliver each of the eight key steps of the evidence synthesis process for systematic / rapid reviews or systematic maps in fields of ecology, environmental management, and conservation.

These frameworks are in part the result of the uptake of standards and protocols were first developed in the field of healthcare and medical research, and in part the product of the development of new *ad hoc* evidence-synthesis to address environmental / conservation issues.

2.3.1 Collaboration for Environmental Evidence (CEE) framework[9]

The Collaboration for Environmental Evidence framework (Collaboration for Environmental Evidence, 2022) is an adaptation of methodologies used in health sciences. Detailed guidelines are available on the CEE website (and synthesis in Appendix 1), along with a range of tools and templates to guide the users and facilitate the preparation of a review up to CEE standards. See also Appendix 1 for more details about this framework.

Applicability

Systematic reviews, rapid reviews, systematic maps.

Assessment against the key steps

1. Planning the synthesis and developing the question. Detailed guidance is provided on how to formulate the question driving the synthesis, how to choose the appropriate method, how to assemble the review team, and how to estimate resourcing requirements and timelines.

2. Developing a protocol. Detailed guidance is provided on how to develop a protocol *a priori* (before the synthesis is conducted) and on the importance of this step to reduce the risk of bias and to ensure that the synthesis is transparent, defensible, and reproducible. CCE provides protocol templates[10,11] as well as the option to register an evidence synthesis protocol in PROCEED, an open access registry of titles and protocols for prospective evidence syntheses in the environmental sector. This is not mandatory but is considered important to avoid duplication of effort and to reduce risk of bias in the conduct of reviews by encouraging the practice of protocol development[12]. The registration of a protocol in PROCEED is free and includes feedback from the editors of the portal.

3. Conducting a search. Detailed guidance is provided on how to conduct a search and on the main sources of bias which may affect the search outcome.

4. Eligibility screening. Detailed guidance is provided on how to set eligibility criteria and carry out the screening process. Study design should be included among the eligibility criteria. The CEE method does not mandate the exclusion of any study design but, depending on the scope of the review and on the time / resources available, the user may decide to focus primarily on

[9] https://environmentalevidence.org/

[10] https://environmentalevidencejournal.biomedcentral.com/submission-guidelines/preparing-your-manuscript/systematic-review-protocol

[11] https://environmentalevidencejournal.biomedcentral.com/submission-guidelines/preparing-your-manuscript/systematic-map-protocol

[12] https://environmentalevidence.org/proceed/

study designs that minimise bias and produce results suitable for inclusion in a meta-analysis as part of step 7 below (data synthesis).

5. Data coding and extraction. Detailed guidance is provided.

6. Critical appraisal of the eligible resources. Detailed guidance is provided for the assessment of internal validity (i.e., the extent of bias in the results of an individual study due to flaws in study design or conduct) and external validity (i.e., the extent to which the results of an individual study can be generalised and applied to other circumstances) of the evidence sources. The use of one or several risk of bias tools to guide the assessment is recommended, but no guidance is provided on this aspect.

The threshold for eligibility is determined by the scores assigned to eight criteria of internal validity and two criteria of external validity. With a simple High vs. Low validity scoring systems, all 10 criteria must receive of High score for the study to be retained in the review. More complex scoring systems can be used, but these should always be categorical (CEE considers that numeric scores may lead to a misleading account of risk of bias).

Frampton et al. (2022) is provided as a key reference for more guidance of risk of bias tools and scoring systems for risk of bias classification.

7. Data synthesis. Detailed guidance is provided for the narrative and quantitative synthesis (based on meta-analysis techniques) of the results. Guidelines for presenting the results of systematic maps are only high-level, although the principles of the narrative synthesis apply to systematic maps. Overall, although the guidance is detailed, browsing through a few CEE systematic reviews / maps in the CEE Library[13] is helpful to better understand how to present the results.

8. Interpreting findings and reporting. Detailed guidance for reporting is provided, and CEE has developed standard reporting formats for systematic reviews and maps[14,15] as well as Reporting standards for Systematic Evidence Syntheses (ROSES[16]), which provide a reporting framework to ensure that evidence syntheses report their methods to the highest possible standards.

The CEE framework does not include an approach to assess the strength of the whole body of evidence assessed in a review. CEE recommends that the review authors openly acknowledge the weakness associated with the validity of the resources examined, the size and statistical significance of the observed effect, the consistency of the effects across studies, the presence of biases and confounding variables, and the clarity of the relationship between the intensity of the exposure / impact and the outcome. However, CEE considers that the overall impact of these limitations on the conclusions of the study can only be considered subjectively and not assessed using a scoring system.

[13] https://environmentalevidence.org/completed-reviews/

[14] https://environmentalevidencejournal.biomedcentral.com/submission-guidelines/preparing-your-manuscript/systematic-review

[15] https://environmentalevidencejournal.biomedcentral.com/submission-guidelines/preparing-your-manuscript/systematic-map

[16] https://environmentalevidence.org/roses/

Pros

- Very detailed guidance for all key steps available on the CEE website[17,18].
- Highlights the need for stakeholder / expert consultation at the planning and protocol development stages.
- Places strong emphasis on the need for a protocol and on the importance of working in teams to reduce errors and biases.
- Places great emphasis on carefully documenting all aspects of the synthesis process, so that the review is transparent, reproducible and defensible.
- It is a great didactic tool, with detailed explanations and a range of templates and tools to guide and assist the user at various stages of the review process. We have provided a condensed version of the CEE guidelines in Appendix 1.
- It is not prescriptive about the exclusion of certain study designs and does not recommend specific risk of bias tools. This would suit skilled users able to assess the risk of bias of different study designs and familiar with a variety of risk of bias tools.

Cons

- Does not provide an approach for assessing the strength of a whole body of evidence.
- The guidelines for rapid reviews are only high-level.
- It is not prescriptive on the exclusion of certain study designs and on the use of a specific risk of bias tools. This may be confusing for users requiring more detailed guidance on the risk of bias of different study designs and on risk of bias tools.

2.3.2 Conservation Evidence (CE) framework[19]

The CE framework has been developed to assess the impact of conservation interventions, through a combination of systematic mapping and expert assessment of synopses of the evidence (Dicks, Hodge, et al., 2014; Sutherland & Wordley, 2018; Sutherland et al., 2020).

Applicability

Systematic maps.

Assessment against the key steps

1. Planning the synthesis and developing the question. The processes followed to develop the question driving the synthesis is clearly explained on the website and in the individual

[17] https://environmentalevidence.org/information-for-authors/guidelines-for-authors/

[18] https://environmentalevidence.org/standards-table/

[19] https://www.conservationevidence.com/

synopses[20]. The synopses also provide good background information about the topic being examined.

2. Developing a protocol. The need for a protocol is outlined on the website. Protocols are registered on the Open Science Framework[21]. The content of the protocol is included in the final synopsis of evidence.

3. Conducting a search. Details about the search process are outlined in the protocols.

4. Eligibility screening. Eligibility criteria are clearly outlined on the website and in the protocols. Only studies that have quantitatively monitored the effect of a conservation action implemented by humans are included in a synopsis.

5. Data coding and extraction. Details provided in the protocols.

6. Critical appraisal of the eligible resources. CE does not quantitatively assess the evidence from each publication or weigh it according to quality, although studies with very evident flaws in the sampling design and in the statistical analyses are immediately discarded. To facilitate the interpretation of the evidence, the size and design of each study is clearly reported in the synopsis.

The strength of the evidence is assessed as a whole by a panel of experts who review the final synopsis. A Delphi method, consisting in several rounds of anonymous scoring and commenting, is used to prevent panellists from influencing each other. The experts score the effectiveness / harm of conservation actions and the certainty of the evidence on a 0-100% scale. A median score across all experts is then calculated for these three parameters (Effectiveness-Harm-Certainty) to determine the Overall Effectiveness of the conservation action, which is expressed on a six-point scale. Each of the six classes of Overall Effectiveness is defined by a series of thresholds for Effectiveness, Certainty and Harm of a conservation action. The Certainty scores provides an assessment of the quality of the evidence, with a score <40% indicating that the quality and / or quantity of the evidence is insufficient to draw any conclusions (Figure 5).

[20] https://www.conservationevidence.com/synopsis/index

[21] https://osf.io/mz5rx/

Overall effectiveness categories		Effectiveness scores	Certainty scores	Harm scores
Beneficial		>60%	>60%	<20%
Likely to be beneficial	Criteria 1	>60%	40% to 60%	<20%
	OR			
	Criteria 2	40% to 60%	≥40%	<20%
Trade-offs between benefits & harms		≥40%	≥40%	≥20%
Unknown effectiveness		Any score	<40%	Any score
Unlikely to be beneficial		<40%	>60%	Any score
Likely to be harmful	Criteria 1	<40%	>60%	Any score
	OR			
	Criteria 2	<40%	≥40%	≥20%

Figure 5: CE framework to assess the effectiveness of conservation actions.

7. Data synthesis. The synthesis of the data is only narrative. No quantitative analyses are carried out, which is typical of systematic maps. A pre-defined format (outlined in the protocol) is used to describe the key features and finding of a study or group of studies (Figure 6).

A [TYPE OF STUDY] in [YEARS X-Y] in [HOW MANY SITES] in/of [HABITAT] in [REGION and COUNTRY] [REFERENCE] found that [INTERVENTION] [SUMMARY OF ALL KEY RESULTS] for [SPECIES/HABITAT TYPE]. [DETAILS OF KEY RESULTS, INCLUDING DATA]. In addition, [EXTRA RESULTS, IMPLEMENTATION OPTIONS, CONFLICTING RESULTS]. The [DETAILS OF EXPERIMENTAL DESIGN, INTERVENTION METHODS and KEY DETAILS OF SITE CONTEXT]. Data was collected in [DETAILS OF SAMPLING METHODS].

A replicated study in 1999–2004 in a wetland on an island in Catalonia, Spain (1) found that all 69 bat boxes of two different designs were used by soprano pipistrelles *Pipistrellus pygmaeus* with an average occupancy rate of 71%. During at least one of the four breeding seasons recorded, 96% of boxes were occupied and occupation rates by females with pups increased from 15% in 2000 to 53% in 2003. Bat box preferences were detected in the breeding season only, with higher abundance in east-facing bat boxes (average 22 bats/box) compared to west-facing boxes (12 bats/box), boxes with double compartments (average 25 bats/box) compared to single compartments (12 bats/box) and boxes placed on posts (average 18 bats/box) and houses (average 12 bats/box). Abundance was low in bat boxes on trees (average 2 bats/box). A total of 69 wooden bat boxes (10 cm deep x 19 cm wide x 20 cm high) of two types (44 single and 25 double compartment) were placed on three supports (10 trees, 29 buildings and 30 electricity posts) facing east and west. From July 2000 to February 2004, the boxes were checked on 16 occasions. Bats were counted in boxes or upon emergence when numbers were too numerous to count within the box.

Figure 6: CE narrative synthesis format.

8. Interpreting findings and reporting. The synopses simply list the narrative synthesis of individual studies or groups of studies, with no additional commentary or interpretation. The expert assessment of the evidence is provided in *What Works in Conservation*[22] (Sutherland et

[22] https://www.conservationevidence.com/content/page/79

al., 2020). Here the key findings of the synopsis are combined with the results of the expert assessment of the evidence, but no additional commentary or interpretation is provided.

Pros

- Provides a well-established framework to tackle broader questions (i.e., lacking the level of detail and focus required by systematic reviews).
- The methods for evidence search and synopsis preparation are transparent and repeatable. The protocols are available online and the synopses on the CE website are constantly updated.
- Provides a clear and repeatable format to prepare synopses of evidence.
- Provides an approach to assess the strength of a whole body of evidence based on thresholds.
- Provides an approach to evidence appraisal based on expert panels including a system to minimise bias in decisions made by groups (i.e., the Delphi method). This multilayered approach is common in medical practice and is supposed to facilitate evidence-informed decision-making (Dicks, Hodge, et al., 2014; Dicks, Walsh, et al., 2014; Walsh et al., 2015).

Cons

- All methods are presented transparently, but with less emphasis on educating / supporting the reader compared to CEE.
- It does not provide guidance on the importance of working in teams to reduce errors and biases.
- There is no appraisal of individual sources of evidence (but this is typical of systematic maps and information about the design of each study is provided).
- The appraisal of the whole body of evidence relies on a well assembled expert panel. No information is provided on the criteria followed for establishing the panels.
- The thresholds used to appraise the whole body of evidence are arbitrary and CE does not explain how they were established.

2.3.3 Eco Evidence framework[23]

The Eco Evidence method is a weight of evidence (WoE) framework designed to assess cause-effect relationships. The framework is supported by an online database of evidence extracted from publications as part of previous studies, and by software guiding the user through the review process. These are available online[24] along with a method manual (Nichols et al., 2011)[25].

[23] https://toolkit.ewater.org.au/Tools/Eco-Evidence

[24] www.toolkit.net.au

[25] https://toolkit.ewater.org.au/Tools/DownloadDocumentation.aspx?id=1000301

Eco Evidence is a product for analysing scientific evidence and assessing its strength and quality. Membership is required to use the tool, and it is only available to those with an Australian address.

Applicability

Systematic reviews, rapid reviews.

Assessment against the key steps

1. Planning the synthesis and developing the question. The method manual provides detailed guidance for the formulation of the question driving the synthesis.

2. Developing a protocol. Eco Evidence does not require the preparation of a formal protocol, but the user is reminded of the importance of carefully documenting all methods used and decisions made.

3. Conducting a search. Very little guidance is provided on how to conduct a search.

4. Eligibility screening. High-level guidance is provided on how to set eligibility criteria and carry out the screening process. Study design is a key criterion for eligibility.

5. Data coding and extraction. This step is facilitated by the Eco Evidence software. Alternatively, a spreadsheet can be used. Detailed guidance is provided on key aspects of the evidence sources that need to be documented (e.g., study design, replication, statistical analyses). The method manual also includes a detailed list of the study designs that can be assessed using Eco Evidence. According to the method manual, this step may require approx. 1h per resource, although there is likely to be large variability among individual reviewers.

6. Critical appraisal of the eligible resources. Eco Evidence assigns a weight score to each resource (resources of higher quality are given higher weight). The weight score is based on: 1) study design; 2) number of independent sampling units used as controls; 3) number of independent sampling units used to investigate impacts (Figure 7).

Study design type	Weight
After impact only	1
Reference/Control vs. impact (no before)	2
Before vs. after (no reference/control)	2
Gradient response model	3
BACI or BARI MBACI or Beyond MBACI	4

Number of control/reference locations	Weight
0	0
1	2
≥ 2	3

Number of impact locations	Weight
1	0
2	2
> 2	3

Number of impact locations	Weight
3	0
4	2
5	4
≥ 6	6

Figure 7: *Weight scores assigned by Eco Evidence to different study designs (see details in the method manual) and different levels of replication for control and impact locations.*

The methods manual provides a transparent justification of the rationale underlying the scoring systems. Users can customise the weight scores, in which case they are reminded of the importance of providing a clear justification.

For each resource, the weight scores for study design and control / impact replication are summed to give an overall study weight ranging between 1 and 10.

7. Data synthesis. The weights of the individual resources are summed to produce the overall weight of the evidence, which is assessed against the thresholds for three causal criteria explaining the relationship between the effect / impact and the response being investigated.

- Criterion 1 and 2: Response and Dose-Response. If the combined weight scores of the studies showing an effect / impact is >20, this is considered a High level of support for the relationship between the effect / impact and the response being investigated (while scores <20 indicate a Low level of support).
- Criterion 3: Consistency. If the combined weight scores of the studies not showing an effect / impact is >20, this is considered a Low level of consistency in the relationship between the effect / impact and the response being investigated (while scores <20 indicate High level of evidence consistency).

These calculations are done by the Eco Evidence software or can be done in a spreadsheet.

Finally, the High / Low scores for the three criteria are combined to determine the final outcome of the assessment as outlined in Table 2.

Table 2: Possible outcome of the assessment of the evidence synthesis based on the combination of scores for the three causal criteria explaining the relationship between the effect / impact and the response being investigated.

Outcomes	Response	Dose-Response	Consistency of Association	Conclusion
Outcome 1	H	H	H	Support for hypothesis
Outcome 2	H	L	H	Support for hypothesis
Outcome 3	L	L	H	Insufficient evidence
Outcome 4	H	H	L	Inconsistent evidence
Outcome 5	H	L	L	Inconsistent evidence
Outcome 6	L	L	L	Support for alternative hypothesis
Outcome 7	No evidence supplied	No evidence supplied	No evidence supplied	No evidence supplied

8. Interpreting findings and reporting. Some high-level guidance to interpret the results of the synthesis is provided. The software produces a full report of the analysis, which details how the evidence used in the assessment was weighted and interpreted.

Pros

- Provides a software guiding the user through the review process.
- Straightforward appraisal on the evidence based on the study design.
- Provides an approach to assess the strength of a whole body of evidence based on thresholds. A justification for the thresholds is provided, and the user is reminded not to apply them unthinkingly.

- Weight scores and thresholds can be customised (this requires careful consideration and documentation).
- Encourages careful documentation of all steps and decisions made.
- It is a good didactic tool, with detailed explanations for many steps of the process in the user manual.

Cons

- Membership is required to use the tool, and it is only available to those with an Australian address.
- Does not require a protocol.
- Does not provide guidance on the importance of working in teams to reduce errors and biases.
- Does not provide guidance on how to shorten a systematic review into a rapid review.
- Provides limited guidance on the search strategy.
- The appraisal approach is simplistic because a study with a strong design can still be affected by other forms of bias.
- Uses numeric scores, while many authors and guidelines (including those of the CEE and US EPA frameworks presented in this report) warn against doing so.

2.3.4 US Environmental Protection Agency (EPA)framework[26]

The United States Environmental Protection Agency (US EPA) has developed a weight of evidence (WoE) framework for environmental assessments. It is used in a variety of circumstances, for example to assess heterogeneous evidence and to determine the causes of observed effects, the hazards posed by chemicals or other agents, and the effectiveness of remediation (Environmental Protection Agency, 2016; Suter et al., 2017).

Applicability

Systematic reviews, rapid reviews.

Assessment against the key steps

1. Planning the synthesis and developing the question. Detailed guideline is provided about key aspects to consider during the planning phase (including quality assurance procedures) and on how to formulate the question driving the synthesis.

2. Developing a protocol. An analysis plan, documenting *a-priori* what will be done, is required, although detailed guidance for its preparation is not provided.

[26] https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100SFXR.TXT

3. Conducting a search. The need for a systematic literature search is clearly explained. Detailed guidelines are not provided, but there is refence to the guidance provided by other platforms (e.g., CEE).

4. Eligibility screening. High-level guidance is provided on how to set eligibility criteria and carry out the screening process. This includes the recommendation to use pairs of screeners to minimise errors.

5. Data coding and extraction. Guidance for data coding / extraction is provided. As part of this process, sources of evidence should be categorised into distinct categories (for example, field vs. laboratory studies). Distinct categories of evidence can be assessed separately before they are integrated to reach a conclusion.

6. Critical appraisal of the eligible resources. The appraisal of the evidence is based on a qualitative weight of evidence scoring system (Figure 8). Symbols are preferred to numerical scores because their use implies that they cannot be numerically combined. Two strongly supporting laboratory tests (++ and ++) are not equal to four somewhat supporting field tests (+, +, +, +). The complexity of the scoring system can be adapted to the assessment and the desired degree of discrimination by increasing or reducing the number of the levels of weight (i.e., by having fewer or more +/- symbols).

+ + +, − − −	Convincingly supports or weakens
+ +, − −	Strongly supports or weakens
+, −	Somewhat supports or weakens
0	No effect (neutral or ambiguous)
NE	No evidence

Figure 8. US EPA qualitative weight of evidence scoring system to represent evidence that, respectively, supports, weakens, or has no effect on the credibility of a hypothesis.

Three properties of each source of evidence (relevance, strength and reliability are scored) and then tabulated. The criteria to be used to score these relevance, strength and reliability are discussed at high-level without any reference to risk of bias tools.

The weights for the three properties are combined into an overall weight for each source of evidence (Table 3). The score for the overall weight is not the average of the property scores. In Table 3, the overall score is only "+" (i.e., somewhat supports) despite moderate strength and high reliability because properties with low weight (relevance in this case) have greater influence than moderate- or high-weight properties.

Table 3: Scoring table with an example of qualitative scores for relevance, strength and reliability combined to determine the overall weight of a source of evidence supporting a certain hypothesis.

Relevance	Strength	Reliability	Overall Weight
+	++	+++	+

7. Data synthesis. The weights of the individual resources are combined to produce the overall weight of a body of evidence. This is done using a weight-of-evidence table. WoE tables can have different structures. Individual evidence sources may carry over directly from the scoring table (Table 3) or may be combined into categories based on study types (e.g., laboratory and field studies) or on the causation mechanism assessed. Ultimately, the scores of sources / categories of evidence are combined to determine the overall weight of the evidence in support of alternative hypotheses.

8. Interpreting findings and reporting. Guidance on how to interpret the results of the synthesis and how to deal with ambiguous results is provided. The interpretation of the evidence requires logic and background knowledge, it is not simply the result of the tabulation of the scores. Brief guidelines on how to present and justify the results are also included.

Pros

- Allows assessment of heterogenous evidence, with no discrimination based on study design or other technical features.
- Simple appraisal system.
- The qualitative scoring systems avoids potential biases introduced by assigning numeric scores.

Cons

- Does not explicitly account for study design and methodology of individual sources of evidence and therefore provides limited guidance on the bias / reliability of the evidence to guide the reviewer.
- Lack of screening based on study design / methodology and loose screening criteria are likely to result in large amounts of resources to be appraised.
- The criteria to be used to score the relevance, strength and reliability are discussed at high-level without any reference to risk of bias tools.
- The method for weighing a whole body of evidence (i.e., step 7 - data synthesis) is not clearly explained.

2.3.5 Comparison of the frameworks

Table 4 below provides a ‘side by side’ comparison of the four frameworks examined above. Table 5 assesses the extent to which each framework is consistent with the criteria outlined in Section 1.4.1 of this report.

RECOMMENDATION 4. KEY FEATURES OF THE FRAMEWORKS

1) The CEE framework is the most detailed and represents a “gold standard”. While it may not be possible for the Ministry to follow this framework to the letter, we recommend that Ministry staff familiarise themselves with the guidance provided by CEE and with some of the other resources referenced therein.
2) A good understanding of the CEE framework and guidance will provide Ministry staff with a good understanding of what the ideal review looks like and where / how adjustments and compromises can be made to speed up the process.
3) The CE framework is an interesting combination of systematic mapping and expert assessment of the evidence which could be adapted to the needs of the Ministry by modifying the expert panel scoring system to answer a broader range of questions (it is only conceived to assess the effectiveness of conservation actions). However, it is important to keep in mind that, with this framework, the ultimate outcome of the synthesis (i.e., judgements on the effectiveness of conservation actions) is based entirely on expert opinion as there is no appraisal of individual evidence sources.
4) The EcoEvidence and US EPA frameworks can appear more user-friendly at first sight because they use straightforward appraisal systems and provide only high-level guidance for some of the evidence synthesis steps. However, it is important to consider that simple appraisal approaches may be too simplistic and may not provide the reviewer with enough guidance. We also recommend referring back to the CEE guidelines when other frameworks do not provide sufficient detail and guidance.

This recommendation supports:

- ☑ Providing for an ‘absolute’ assessment of evidence
- ☑ Transferability of a process across policy questions in different environmental domains and at different levels of government – there remains flexibility as to *how* each of the steps is delivered to adapt to the different circumstances.
- ☑ Implementation of a repeatable process that allows for follow up assessments in the future.
- ☑ Applicability to different stages of the policy cycle
- ☑ A transparent evaluation of scientific evidence – with the extent to which the likelihood of selection biases is reduced dependent on *how* certain steps are undertaken.

Table 4: Comparison of how the assessed methods achieve the key steps in an evidence synthesis process.

Framework	CEE	CE	Eco Evidence	US EPA
Applicability	**Systematic reviews**	**-**	**Systematic reviews**	**Systematic reviews**
	Rapid reviews	**-**	**Rapid reviews**	**Rapid reviews**
	Systematic maps	**Systematic maps**	**-**	
1. Planning the review and formulating the question	Detailed guidance on how to formulate the question driving the synthesis, how to choose the appropriate method, how to assemble the review team, and how to estimate resourcing requirements and timelines.	The processes used to develop the question driving the synthesis is clearly explained on CE's website and in the individual synopses. No guidance on how to plan the review (e.g., assembling the team, estimating resourcing requirements, etc.)	Detailed guidance on how to formulate the question driving the synthesis. No guidance on how to plan the review (e.g., assembling the team, estimating resourcing requirements, etc.)	Detailed guideline is on key aspects to consider during the planning phase (including quality assurance procedures) and on how to formulate the question driving the synthesis.
2. Protocol	Detailed guidance how to develop a protocol and its importance.	Protocols for CE syntheses protocols are registered on the Open Science Framework and included in the final synopses of the evidence.	A protocol is not required, but emphasis is placed on the importance of carefully documenting all methods used and decisions made.	An analysis plan, documenting *a-priori* what will be done, is required, but detailed guidance for its preparation is not provided.
3. Search	Detailed guidance on how to conduct a search and on the main sources of bias which may affect the search outcome.	The protocol includes details about the search process.	Limited guidance provided.	Limited guidance provided, but the need for a systematic literature search is clearly explained.
4. Screening	Detailed guidance on how to set eligibility criteria and carry out the screening process. Study design should be included among the eligibility criteria, but the method does not mandate the exclusion of any study design.	CE only includes studies that have quantitatively monitored the effect of a conservation action implemented by humans. Screening criteria are reported in the protocol.	High level guidance on how to set eligibility criteria and carry out the screening process. Study design is a key criterion for eligibility.	High-level guidance provided. Loose screening with no discrimination based on study design and methodology.
5. Data coding and extraction	Detailed guidance provided.	Detailed in the protocol.	Detailed guidance on key aspects of evidence sources that need to be documented (e.g., study design, replication, statistical analyses).	High-level guidance provided.
6. Appraisal	Detailed guidance for the assessment of internal and external validity of the evidence sources. The use of one or several risk of bias tools is recommended, but no guidance is provided on this aspect.	No appraisal of individual evidence sources, but studies with evident flaws are immediately discarded. The strength of the evidence is assessed as a whole by a panel of experts who review the final synopsis. The expert assessment produces a score based on a threshold system.	The appraisal of the evidence is based on a numeric weight of evidence (WoE) scoring system. Scores are based on study design and replication.	The appraisal of the evidence is based on a qualitative weight of evidence (WoE) scoring system based on +/- operators. Scores are based on study relevance, strength and reliability. The criteria to be used to score these relevance, strength and reliability are discussed at high-level without any reference to risk of bias tools.

Framework	CEE	CE	Eco Evidence	US EPA
7. Data synthesis	Detailed guidance on narrative and quantitative synthesis (based on meta-analysis techniques) of the results. Guidelines for presenting the results of systematic maps are only high-level.	Narrative synthesis based on a pre-defined format.	WoE scores of the individual resources are summed to produce an overall weight of the evidence. This is assessed against thresholds for causal criteria explaining the relationship between the effect/impact and the response being investigated.	The weights of the individual resources are combined to produce the overall weight of a body of evidence. This step is not described very clearly.
8. Reporting	Detailed guidance and reporting templates are provided. The CEE framework does not include an approach to assess the strength of the whole body of evidence assessed in a review.	The synopses list the narrative synthesis of the individual studies, without any additional commentary or interpretation. The expert panel assessment of the evidence is provided in *What Works in Conservation* (Sutherland et al., 2020).	High-level guidance provided.	Guidance on how to interpret the results of the synthesis and how to deal with ambiguous results is provided.
Summary	Very detailed guidance, making it a great didactic tool and a landmark in the field of environmental evidence syntheses. Places great emphasis on transparency, on the need for a protocol and on the importance of working in teams. Provides good guidance for the planning phases preceding the technical work. Guidance on rapid reviews and risk of bias tools is only high-level. Does not provide an approach for assessing the strength of a whole body of evidence.	Well developed and detailed framework for systematic maps. Provides a clear and repeatable format to prepare synopses of evidence. Provides an approach to assess the strength of a whole body of evidence based on expert judgment and a threshold system. No explanation on how the thresholds were developed. No appraisal of individual sources of evidence, therefore the ultimate outcome of the synthesis (i.e., judgements on the effectiveness of conservation actions) is based entirely on expert opinions.	Detailed framework based on weight of evidence scores combined with a threshold system. Clear justification for the scores and thresholds is provided and they can be customised. Provides a software guiding the user through the review process, but it is only available to Australian users. Does not require a protocol but encourages careful documentation of all steps. Limited guidance on search strategies. Straightforward but simplistic appraisal of the evidence. No guidance on how to shorten a systematic review into a rapid review. Uses numeric scores despite widespread opposition to this approach.	Weight of evidence framework allowing the assessment of heterogenous evidence, with no discrimination based on study design or other technical features. The qualitative scoring systems avoids potential biases introduced by assigning numeric scores. Provides limited guidance on the bias / reliability of the evidence to guide the reviewer. Loose screening criteria are likely to result in large amounts of resources to be appraised. The method for weighing a whole body of evidence is not clearly explained.

Table 5: Assessment of the various methods against the characteristics of a potential system

Framework	CEE	CE	Eco Evidence	US EPA
Applicability	**Systematic reviews**		**Systematic reviews**	**Systematic reviews**
	Rapid reviews		**Rapid reviews**	**Rapid reviews**
	Systematic maps	**Systematic maps**		
Allows for an 'absolute' assessment of evidence	Yes	Yes	Yes	Yes
Is transferable across policy questions in different environmental domains and at different levels of government	Yes, although as a process that is relatively resource and time intensive (and is best undertaken by people with specialist expertise in evidence synthesis as well as subject matter experts), it may be less accessible to smaller or less well-resourced government organisations.	Yes, but the expert panel scoring system would need to be modified to be applicable to a broader range of questions (it is only conceived to assess the effectiveness of conservation actions). As a process that is relatively resource and time intensive (and is best undertaken by people with specialist expertise in evidence synthesis as well as subject matter experts), it may be less accessible to smaller or less well-resourced government organisations.	Yes, Eco Evidence is designed to be used by anyone required to review literature on a specific topic of interest, targeting in particular researchers and students in ecology, environmental policy makers and practitioners in river and stream restoration/conservation. It may therefore be a method that is well suited to levels of government that are less well-resourced.	Yes. This framework has been used in a variety of circumstances, for examples to assess heterogeneous evidence and to determine the causes of observed effects, the hazards posed by chemicals or other agents, and the effectiveness of remediation.
Is repeatable and allows for follow up assessments in the future	Yes, if all actions and decisions are thoroughly documented.	Yes, if all actions and decisions are thoroughly documented. The CE synopses are constantly updated.	Yes, if all actions and decisions are thoroughly documented. It does not require the preparation of a protocol.	Yes, if all actions and decisions are thoroughly documented. It does not provide detailed guidance on the preparation of a protocol.
Can be applied to different stages of the policy cycle	The CEE method has potential applicability to all stages of the policy cycle but given the time and resources that are generally required to be invested in this approach (particularly when combined with the potentially short timeframes of the political cycle in New Zealand) it is likely to be best suited to the decision-making stage and the evaluation stage.	Similar to the CEE method, if the method that is used by CE were to be used to undertake evidence synthesis to inform various stages of the policy cycle, it would likely be best suited to the decision-making stage and the evaluation stage.	Similar to the CEE method, if the method that is used by Eco Evidence were to be used to undertake evidence synthesis to inform various stages of the policy cycle, it would likely be best suited to the decision-making stage and the evaluation stage due to the level of time and resources that are required to be invested in this approach. However, if time and resources permit, it could be applied to the policy formulation stage as well.	Similar to the CEE method, if the method that is used by USEPA were to be used to undertake evidence synthesis to inform various stages of the policy cycle, it would likely be best suited to the decision-making stage and the evaluation stage due to the level of time and resources that are required to be invested in this approach. However, if time and resources permit, it could be applied to the policy formulation stage as well.
Ensures a transparent evaluation of scientific evidence and reduces the likelihood of selection biases	Yes	Yes, but it makes use of a "custom-made" system of scores and thresholds which does not seem to have had any further uptake so far. The ultimate outcome of the synthesis is based entirely on expert judgment.	Yes, but it makes use of a "custom-made" system of scores and thresholds which does not seem to have had any further uptake so far.	Yes, but some aspects of the appraisal and synthesis of the evidence are not clearly explained.

Framework	CEE	CE	Eco Evidence	US EPA
Time requirements[27]	Very detailed and time-consuming protocols. For a team of three people: • Approx. 12 months for a systematic review / map • Up to 6 months for a rapid review, but this could be reduced further imposing significant restrictions for some of the steps (e.g., setting restrictive search and eligibility criteria).	Very detailed and time-consuming protocol. Approx. 12 months for a team of three people	It depends on the searching criteria established by the user (the framework is not prescriptive in this regard). The methods for screening, appraisal and synthesis (intuitive and easy to follow) make this method well suited for rapid reviews.	It depends on the searching criteria established by the user (the framework is not prescriptive in this regard). Loose screening criteria are likely to result in large amounts of resources to be appraised.

[27] An online tool is available to estimate how long a review will take to complete: https://predicter.github.io/#tool. The tool is described in Haddaway & Westgate (2019).

2.4 Other tools for evidence synthesis

In addition to the frameworks described above, there is a range of additional tools which are not part of a framework but can be incorporated into an evidence synthesis. The following sections provide an overview of some of the tools developed for the fields of ecology, environmental science, and conservation and of other tools which are widely used in other fields.

2.4.1 Appraisal tools for individual studies

Cochrane RoB 2[28]

The risk-of-bias tool for randomised trials (RoB 2) is Cochrane's recommended tool to assess the risk of bias in randomised trials. RoB 2 is structured into five domains through which bias might be introduced into the result: 1) bias arising from the randomization process; 2) bias due to deviations from intended interventions; 3) bias due to missing outcome data; 4) bias in measurement of the outcome; 5) bias in selection of the reported result.

For each domain, there is a checklist of signalling questions (three to seven questions per domain) guiding the reviewer in the assessment of the features of the study that are vulnerable to risk of bias. The tool includes algorithms that convert the responses to the signalling questions into risk-of-bias judgement for each of the five domains, which is expressed on a three-point scale:

- Low risk of bias
- Some concerns
- High risk of bias

A judgement of "High risk of bias" for any individual domain will lead to the study being considered as at "High risk of bias" overall. In the absence of high risk of bias, a judgement of "Some concerns" for any individual domain will lead to an overall "Some concerns" categorization for the study. In the absence of judgments of both "High risk of bias" and "Some concerns", the study is judged to be at low risk of bias.

Cochrane ROBINS-I[29]

The ROBINS-I tool (Risk Of Bias In Non-randomised Studies of Interventions) is Cochrane's recommended tool to assess the risk of bias in non-randomised studies in which the subjects of the study are allocated to the intervention and the control group in a non-random fashion. ROBINS-I is structured into seven domains through which bias might be introduced into the result. The seven domains encompass two forms of pre-intervention biases, one form of bias during the intervention, and four forms of post-intervention bias.

For each domain, there is a checklist of signalling questions guiding the reviewer in the assessment of the risk of bias. If none of the answers to the signalling questions for a domain suggest a potential problem, then risk of bias for the domain can be judged to be low. Otherwise, potential for bias exists and the reviewer must make a judgement on the extent to

[28] https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool?authuser=0

[29] https://www.riskofbias.info/welcome/home/current-version-of-robins-i

which the results of the study are at risk of bias. Risk-of-bias judgement for each of the seven domains is expressed on a five-point scale:

- Low risk of bias - the study is comparable to a well-performed randomised trial with regard to a specific domain.
- Moderate risk of bias - the study is sound for a non-randomised study but cannot be considered comparable to a well-performed randomised trial with regard to a specific domain.
- Serious risk of bias - the study has some important problems with regard to a specific domain.
- Critical risk of bias - the study is too problematic with regard to a specific domain to provide any useful evidence.
- No information on which to base a judgement about risk of bias with regard to a specific domain.

The same five-point scale is then used to express the overall risk of bias judgment for the study:

- Low risk of bias - The study is judged to be at low risk of bias for all domains (the study is comparable to a well-performed randomised trial).
- Moderate risk of bias - The study is judged to be at low or moderate risk of bias for all domains (the study appears to provide sound evidence for a non-randomised study but cannot be considered comparable to a well-performed randomised trial).
- Serious risk of bias - The study is judged to be at serious risk of bias in at least one domain, but not at critical risk of bias in any domain (the study has some important problems).
- Critical risk of bias - The study is judged to be at critical risk of bias in at least one domain (the study is too problematic and should not be used).
- No information on which to base a judgement about the study risk of bias.

Collaboration for Environmental Evidence Critical Appraisal Tool[30]

CEE is developing a critical appraisal tool for studies assessing effectiveness of interventions or impacts of exposures in environmental management. The tool is in draft form and is still being trialled.

The tool is applicable to both experimental and observational studies and does not allow risk of bias to be assessed solely based on the design of the study. The tool is based on seven risk of bias criteria (equivalent to the domains used by RoB 2 and ROBINS-I) and provides checklists to help judgement about risk of bias within each risk-of-bias criterion. Once assessors have responded to all checklist questions within a risk-of-bias criterion, they will have to judge the risk of bias for the criterion. The risk-of-bias for each of the seven criterial is expressed as:

- Low risk of bias
- Medium risk of bias
- High risk of bias

[30] https://environmentalevidence.org/cee-critical-appraisal-tool/

The same categorization is used to express the overall risk of bias of the study:

- Overall low risk of bias - The study is considered to have low risk of bias for all assessed risk-of-bias criteria.
- Overall medium risk of bias – The study is considered to have medium risk of bias for at least one criterion, but there are no criteria with high risk of bias.
- Overall high risk of bias – The study is considered to have high risk of bias for at least one criterion.

For another adaptation of the Cochrane's tools to the field of environmental studies see the Environmental-Risk of Bias Tool proposed by Bilotta et al. (2014).

Balance Evidence Assessment Method (BEAM)

The BEAM is a weight of evidence (WoE) tool that allows the assessment of a diverse range of evidence from a wide variety of sources including local expert and practical knowledge, indigenous and knowledge, studies and syntheses from the scientific literature, and the grey literature across the social and natural science spectrum. No external or a priori hierarchy or system is enforced on pieces of evidence. Instead, all pieces of evidence start out on a level playing field and are scored based on information reliability (I), source reliability (S), and relevance of the piece of evidence (R) to the question being asked. Each piece of evidence receives a score for each of I, S and R (on a scale from 0 to 3 or 0 to 5), which are then multiplied together to produce an overall weight of evidence score (e.g., 3 x 3 x 3 = 27 or 5 x 5 x 5 = 125; Christie et al., 2023; Sutherland, 2022).

Sutherland (2022) provides a series of thresholds to convert the weight of evidence sources into categories of evidence strength (Figure 9) but does explain the rationale underlying the thresholds. It is also worth reiterating that several authors and guidelines disagree with the use of numerical scores to describe the quality of evidence sources (see Sections 2.3.1, 2.3.4, Appendix 1 and Frampton et al., 2022).

Weight	Description of evidence strength
0–1	Unconvincing piece of evidence
2–8	Weak piece of evidence
9–27	Fair piece of evidence
28–64	Reasonable piece of evidence
65–125	Strong piece of evidence

Figure 9: Conversion of weights of single pieces of evidence (obtained by grading I, S, R on a 0-5 scale and by multiplying the three scores) into descriptions of evidence strengths. From Sutherland (2022).

Custom-made tools

By far the majority of environmental science studies do not follow “official” criteria and tools for the appraisal of the evidence but instead develop and use their own. This is an acceptable practice since tools and protocols in the field of environmental science are not as well established as in medical research, but it has been followed by many authors without careful consideration. This has caused the proliferation of an excessive amount of tools of unclear validity and reliability (Frampton et al., 2022; Stanhope & Weinstein, 2023).

Therefore, it is important to beware of tools of unclear origin and to keep in mind that their previous use does not guarantee their validity. Before developing a new appraisal tools or using one of unclear origin, it is recommended to consult with experienced topic experts and to ensure that the tool satisfies the FEAT principles (Frampton et al., 2022):

- FOCUSED: the tool should measure what it claims to measure, i.e., the internal validity of a study.
- EXTENSIVE: the tool should be comprehensive and include all the classes of bias that could arise in a study.
- APPLIED: the tolls should produce an output able to inform the data synthesis stage of the review (for example by dictating the exclusion of the studies with high risk of bias).
- TRANSPARENT: the tool should have a clear rationale, instructions and outputs.

Hierarchies of evidence

There are a number of hierarchies of evidence ranking studies solely based on study design. In these hierarchies, study designs with the lowest risk of bias (e.g., randomised controlled trials) are top-rated and study designs with a higher risk of bias have a lower ranking. While these hierarchies are still often used in systematic reviews, this approach is too simplistic given that there are many other features of a study that can affect its quality along with the design (Bilotta et al., 2014; Stanhope & Weinstein, 2023).

Stanhope & Weinstein (2023) provide a detailed discussion of the limitations of the hierarchies of evidence used for critical appraisal in ecology and ultimately recommend against their use.

RECOMMENDATION 5. SELECTING AN APPRAISAL TOOL FOR INDIVIDUAL STUDIES

1) Consider whether the use of numerical scores in certain tools is appropriate or not. Many authors and guidelines warn against doing so.
2) Beware of tools of unclear origin and do not assume that they have been correctly developed and tested.
3) Before developing a new tool or using one of unclear origin, consult with experienced topic experts and ensure that the tool satisfies the FEAT principles.
4) Do not use hierarchies of evidence only as appraisal tools.

This recommendation supports:

- ☑ Providing for an 'absolute' assessment of evidence
- ☑ Transferability of a process across policy questions in different environmental domains and at different levels of government – there remains flexibility as to *how* each of the steps is delivered to adapt to the different circumstances.
- ☑ Implementation of a repeatable process that allows for follow up assessments in the future.
- ☑ Applicability to different stages of the policy cycle
- ☑ A transparent evaluation of scientific evidence – with the extent to which the likelihood of selection biases is reduced dependent on *how* certain steps are undertaken.

2.4.2 Appraisal tools for reviews

AMSTAR 2[31,32]

AMSTAR is a popular instrument for critically appraising systematic reviews of randomised controlled clinical trials. AMSTAR 2 guides the reviewer through a 16-item checklist available online[33], with some items considered critical and other non-critical. The final output is a qualitative rating of the confidence in the results of the review on a four-point scale:

- High confidence - No or one non-critical weakness. The systematic review provides an accurate and comprehensive summary of the available studies.
- Moderate confidence - More than one non-critical weakness, but no critical flaws. The review may provide an accurate summary of the available studies.
- Low confidence - One critical flaw with or without non-critical weaknesses. The review may not provide an accurate and comprehensive summary of the available studies.
- Critically low confidence - More than one critical flaw with or without non-critical weaknesses. The review should not be relied on to provide an accurate and comprehensive summary of the available studies.

There are examples of applications of AMSTAR (the predecessor of AMSTAR 2) to environmental studies (e.g., Rowland et al., 2021).

[31] https://amstar.ca/index.php

[32] https://www.bmj.com/content/358/bmj.j4008

[33] https://amstar.ca/Amstar_Checklist.php

ROBIS (Risk of Bias in Systematic Reviews)[34] is a similar tool to AMSTAR 2 but considered more difficult to use and better suited to advanced users (Perry et al., 2021).

GRADE[35]

The Grades of Recommendation, Assessment, Development and Evaluation Working Group (GRADE Working Group) has developed a system for grading the certainty of evidence (GRADE) adopted by over 100 organizations worldwide but still relatively untested outside of the fields of healthcare and medical research.

The GRADE approach specifies four levels of certainty for a body of evidence: High, Moderate, Low and Very low. The starting point for rating the certainty of evidence is based on the study design of the resources included in the synthesis. Randomised trials are considered to provide high certainty and non-randomised studies, including observational studies, are considered to provide low certainty.

The assessment of the certainty of the evidence is then refined through consideration of five domains: 1) risk of bias; 2) inconsistency; 3) indirectness; 4) imprecision; 5) publication bias. The certainty score of both syntheses based on randomised and non-randomised studies can be downgraded depending on the presence of these sources of bias. Usually, the certainty rating will fall by one level for each source of bias, up to a maximum of a three-level downgrade. If there are very severe problems for any one domain, evidence may fall by two levels due to that source of bias alone.

Evidence syntheses bases on non-randomised studies (and rarely randomised studies) can instead be upgraded through consideration of three further domains: 1) large effects; 2) evidence of dose-response gradient; 3) all plausible sources of bias would reduce a demonstrated effect or suggest a spurious effect when results show no effect.

CEESAT[36]

CEESAT is the Collaboration for Environmental Evidence Synthesis Assessment Tool for the appraisal of reviews. CEESAT is based on 13 criteria relevant to the evaluation of the objectivity, transparency and comprehensiveness of policy-relevant evidence syntheses in conservation and environmental science. The criteria are based on the key steps of a systematic evidence synthesis (see Section 2.2). For each criterion, scoring guidelines are provided to assist the reviewer. The quality of the review in relation to each criterion is assessed as:

- GREEN (3 points)
- AMBER (1 point)
- RED (0 point)

The total scores of the assessment can range between 0 and 39 and be used to gauge the reliability of a synthesis. However, CEE does not provide thresholds for the interpretation of the total score and considers that the CEESAT scores are better used as a comparative tool when comparing multiple syntheses. In addition to the total score, considering the scores for individual criteria can also be informative to understand the strength and weakness of a review (Woodcock et al., 2014).

[34] https://www.bristol.ac.uk/population-health-sciences/projects/robis/

[35] www.gradeworkinggroup.org

[36] https://environmentalevidence.org/ceeder/about-ceesat/

As reported by Woodcock et al. (2014), CEESAT has been thoroughly evaluated in terms of applicability to different syntheses, validity of scores awarded, effectiveness at discriminating between syntheses, and repeatability.

Evidence Assessment Tool for Ecosystem Services and Conservation Studies[37]

Aside from CEESAT, the Evidence Assessment Tool for Ecosystem Services and Conservation Studies (Mupepele et al., 2016) appears to be the only other tools developed specifically to appraise environmental evidence syntheses. This tool consists of a 24 point checklist, with reviews receiving one point when they satisfy the criterion set out by each point of the checklist. If all 24 criteria are satisfied, the review scores 100% quality points. The quality point score decreases linearly with the number of criteria not satisfied. The scores are then used to rank the reviews as follows:

- 75 – 100% = Very strong evidence
- 50 - 75% = Strong evidence
- 25 – 50% = Moderate evidence
- < 25% = Weak evidence

The rationale underlying the thresholds is not explained and they seem excessively 'permissive': a review could fail to satisfy up to 6 criteria and still be considered very strong evidence.

RECOMMENDATION 6. SELECTING AN APPRAISAL TOOL FOR REVIEWS

1) There are not many tools developed specifically to appraise environmental evidence syntheses. CEESAT appears to be the most comprehensive and reliable tool for this purpose.
2) Tools developed in other fields, like AMSTAR-2, may be applicable to the appraisal of environmental evidence syntheses.

This recommendation supports:

- ☑ Providing for an 'absolute' assessment of evidence
- ☑ Transferability of a process across policy questions in different environmental domains and at different levels of government – there remains flexibility as to *how* each of the steps is delivered to adapt to the different circumstances.
- ☑ Implementation of a repeatable process that allows for follow up assessments in the future.
- ☑ Applicability to different stages of the policy cycle
- ☑ A transparent evaluation of scientific evidence – with the extent to which the likelihood of selection biases is reduced dependent on *how* certain steps are undertaken.

[37] https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/15-0595

2.5 From evidence to decision, communicating the science to policy makers

Environmental policy decisions in New Zealand can range from those made at the central government level (e.g., developing or amending a national policy statement) to plan making at a regional or district council level. This may mean that there will be constraints on the time and resources available to synthesise evidence to inform a decision. Policymakers generally work in fast-paced environments (and under fast-paced processes), are often time-poor, and can be bombarded with information from a range of sources (Wood, n.d.). Collectively, these factors combine to produce a challenging context within which the communication of the findings of evidence synthesis needs to be made to inform a decision.

Given the context within which policy makers often need to make decisions, they are likely to be particularly interested in understanding the following characteristics of the evidence synthesis (in addition to the findings of the synthesis):

- Characteristics of the evidence that informed the synthesis
    - The breadth of findings in the evidence
    - The level of confidence in the findings of the evidence
    - Trends in the evidence
    - Evidence gaps
    - Advice on the policy / decision making implications in the absence of confidence about the success of an intervention
- Understanding the evidence
    - Lay person explanation of the evidence synthesis method used; and/or the methods used in the studies that have informed the synthesis.
- Understanding any constraints within the evidence, or on the delivery of the synthesis, e.g.
    - Whether (and to what extent) the evidence synthesis method deviated from best or good practice (and why).
    - Time and/or resourcing constraints

Accordingly, the findings of evidence syntheses commonly incorporate the following components:

- A high-level summary incorporating:
    - Background
    - Methods
    - Results
    - Conclusions
- Introduction / background to the evidence synthesis, usually capturing:
    - A description of the issue / problem
    - The aim of the evidence synthesis

    - The interventions assessed
    - The outcomes measured
- Method used for the evidence synthesis:
    - Commissioner of the research
    - Involvement of stakeholders in study design
    - Type of evidence synthesis method employed
    - Whether a protocol was prepared
    - Approach to search for studies (including any limitations)
    - Approach to screening of the studies
    - Approach to critical appraisal of the studies and their findings
- Findings and results overall:
    - Findings and results of the evidence in relation to each intervention assessed
- Discussion and conclusions:
    - Trends in the evidence
    - Further research needed
    - Policy considerations

The most effective method to communicate the findings of an evidence synthesis will probably differ depending on the topic, the specific policy audience, nature of the parliamentary or other governmental process, timeframes and current events (Kenny et al., 2018).

The way in which these findings are presented can vary from relatively dense technical papers through to one- or two-page non-technical summaries. Communicating findings in a way that enables policy makers to get a quick overview of the review while also providing links to additional information is likely to be particularly valuable.

Providing information about the findings across the range of interventions assessed in a consistent manner will also assist decision makers to more easily compare and contrast the levels of evidence supporting a range of interventions to achieve a particular outcome.

One example of this approach is set out in Randall & Donnison, 2014, which presents the findings of a systematic map assessing five on-farm interventions to improve water quality. It presents the findings of a systematic map (Randall et al., 2015) in a format that is digestible to policymakers. It features the following:

- A succinct two-page summary at the front of the document describing the synthesis methods used; the interventions assessed, the purpose of the study; a high-level overview of the findings in relation to each intervention; and limitations/further research required.
- Succinct background to the policy problem; the aim of the study; the interventions assessed, and the outcomes measured.
- An overview of the findings in relation to the body of evidence, using flow charts and tables to demonstrate the screening and eligibility assessment processes followed.
- Consistent structure to communicate the findings in the evidence for each intervention:

    - Type of evidence found
    - Variabilities in the evidence
    - Level of scientific rigour in the evidence
    - Limitations of the evidence
    - Effectiveness of the intervention at achieving the stated outcomes
    - Policy implications
    - Research gaps and recommendations for further primary research
    - References of interest
- Use of subheadings, charts, tables and diagrams to support the communication of the findings.

In addition to the above, which highlights the *types* of information arising from an evidence synthesis that are likely to be of particular interest to policy makers, the below outlines supporting *methods* that are described in the literature to support communication at the science-policy interface in a broader sense.

Communication between policy makers and scientific experts at various stages of the process

Good communication between policymakers and scientific experts is required at various stages in the policy-making process, not just when the synthesis of evidence has been completed (Cooke et al., 2023). However, it is also recognised that the intentions of experts and policymakers may often conflict with one another (Kano & Hayashi, 2021). Supporting actors in the policy-making process to become more aware of the integrity of science (independent of the surrounding epistatic values) allows scientists and policy players to align their perspectives in a clearer way (Kano & Hayashi, 2021).

Targeting information to the audience

Good science communication understands and is targeted for the audience receiving the evidence. INASP (2016) explains the types of information necessary at the three decision-making levels: political, strategic, and operating, where political actors make decisions on the orientation of the certain policy, strategic actors are responsible for the policy design, and operating managers are in charge of policy implementation.

Recognising time constraints

Policy decisions may need to be made in circumstances with constraints on time. Providing well-summarised evidence in a timely manner and packaged in a form that meets the needs of practitioners can assist to meet this challenge (Cooke et al., 2023; Walsh et al., 2015). The use of synopses is an effective tool in completing a systematic literature review (Dicks, Hodge, et al., 2014). Cooke et al. (2023) recommend the use of lay language summaries as part of the synthesis process and that lay language text should be traceable (using hyperlinks, for example) to detailed scientific descriptions and original sources. Superu (2018) suggests taking a multi-faceted approach to sharing the information (i.e., written documents as well as presentations / interactive sessions) when sharing the evidence findings, as well as keeping it simple, easy to read, and supported with graphics and images.

Appointing credible scientific advisers

When considering legitimacy, the dilemma between science and democracy cannot be avoided. Lacking the ability to assess scientific expertise, the public and policymakers are required to delegate some aspects of policy decision-making to experts. As such, policymakers need to select credible scientific advisers and collect evidence appropriately in response to policy objectives, thus requiring the partial transfer of decision-making to professional organisations or groups without democratic representation (Kano & Hayashi, 2021). It is important to ensure that those responsible for facilitating knowledge translation have the appropriate specialist skills, with competencies described in European Commission, Joint Research Centre (2017), Gensby et al. (2019), and Miljand & Eckerberg (2022).

RECOMMENDATION 7. COMMUNICATING THE FINDINGS

It is recommended that the following aspects are considered in the overall design, implementation and resourcing of a potential process for assessing the strength of scientific evidence and communicating those findings to policymakers.

1) That the findings of an evidence synthesis are structured in a way that enables a policymaker to quickly identify the findings of the synthesis (e.g., in a short executive summary), while also providing further information and links for more of a 'deep dive' on particular aspects of the synthesis.
2) That the findings of an evidence synthesis include (both in the summary and in the main body of the findings):
    a) The policy problem/question; interventions assessed, and outcomes measured.
    b) The evidence synthesis method followed.
    c) A description of the method (and any limitations arising as a result of time/resource constraints).
    d) The type of evidence found.
    e) Variabilities in the evidence.
    f) Level of scientific rigour in the evidence.
    g) Limitations of the evidence.
    h) Effectiveness of the intervention at achieving the stated outcomes.
    i) Policy implications.
    j) Research gaps and recommendations for further primary research.
    k) References of interest.
3) That the method of communicating the findings of evidence synthesis ideally:
    a) Is multi-faceted (e.g., a combination or selection of documents, presentations and interactive sessions).
    b) Is supported with graphics and images.
    c) Uses lay language that is traceable to detailed scientific descriptions and original sources (e.g., by using hyperlinks).
    d) Is targeted to, and informed by an understanding of, the audience receiving the summaries.
4) That suitably trained science communicators with the appropriate specialist skills are involved as part of the team in delivering an evidence synthesis.
5) That summaries are provided in a form that meets the needs of practitioners – reinforcing the need for science communicators to be involved in planning stages of the synthesis project.

To assist with scenarios where there are significant time / resource constraints, there may be some benefit in sourcing some examples of high-quality summaries of evidence synthesis that could be translated into 'templates'.

These recommendations may assist with improving communication and dissemination skills between scientists and decision-makers.

3. Summary of findings and recommendations for the development of evidence-informed policy in New Zealand

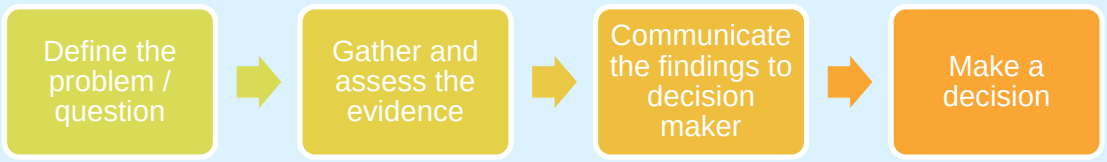
There is significant coherence as to what the principles of evidence-informed policy should be; transparent, credible and legitimate are most frequently proposed across the literature reviewed (Christie et al., 2022; Cooke et al., 2023; Kano & Hayashi, 2021; S. J. Nichols et al., 2017; Sarkki et al., 2014; Schwartz et al., 2018; Superu, 2018; United Nations Environment Programme, 2020). Other principles referred to for good evidence-informed policy-making include robustness, repeatable, accessible and rapid (Cooke et al., 2023; Nichols et al., 2017), the latter speaking to issues of time and resource.

In order to ensure a transparent, credible and legitimate approach, our findings from the literature review have identified the following principal recommendations, as highlighted throughout the report.

RECOMMENDATION 1. FRAMEWORKS USED FOR EVIDENCE-INFORMED POLICY

The use of a framework to guide researchers and policymakers through a set of structured and transparent stages or steps in developing evidence-informed policy is recommended, and that the framework should include the following steps as illustrated below:

1) Define the policy problem / question.
2) Gather and assess the evidence (evidence synthesis).
3) Communicate the findings of the evidence synthesis to the decision-maker.
4) Make a decision.



This recommendation supports:

- ☑ A transferable process
- ☑ A repeatable process
- ☑ A transparent evaluation of scientific evidence, aiming to reduce bias
- ☑ A high-level framework within which the steps undertaken (see Recommendation 7) can be tailored to the time, capacity and resources available.

RECOMMENDATION 2. SELECTING THE EVIDENCE SYNTHESIS METHOD

It is recommended that the following factors are used to inform a decision about the most appropriate method of evidence synthesis to use:

1) The nature of the question or problem.
2) The level of certainty required from the synthesis.
3) The time and resources available.

Simple flow charts (see Figure 4) can help with the choice among different evidence synthesis methods.

This recommendation supports:

- ☑ Using a repeatable process to decide which evidence synthesis method to use
- ☑ Selecting the best method to apply at different stages of the policy cycle (in response to the time, resources and expertise available at each stage)

RECOMMENDATION 3. KEY STEPS IN THE EVIDENCE SYNTHESIS PROCESS

1) The eight-step process for systematic evidence synthesis should always be followed.
2) Preparing and following a detailed protocol throughout all the steps is mandatory and it is what sets solid, systematic syntheses aside from other methods.
3) Subjectivity and technical constraints cannot be entirely removed from the evidence synthesis process. All the limitations of the study and the key decisions made by the review team (for example, for the screening and appraisal of the evidence) should be carefully documented and explained in the protocol and in the final evidence synthesis report.
4) Any measure implemented to reduce time requirements as part of a rapid review should be thoroughly documented and motivated.
5) Any deviation from the original protocol should be thoroughly documented and motivated.
6) The first two steps of the process (covering general planning, formulation of the question and preparation of the protocol) should include consultation with relevant stakeholders (and with subject expert / advisory panels if needed) to inform the goals and structure of the review and to ensure that the process of evidence synthesis is as free from bias as possible.
7) Consider the possibility of adopting some of the methods and principles of systematic reviews when conducting a traditional review.

This recommendation supports:

- ☑ Transferability of a process across policy questions in different environmental domains and at different levels of government – there remains flexibility as to *how* each of the steps is delivered to adapt to the different circumstances.
- ☑ Implementation of a repeatable process that allows for follow up assessments in the future.
- ☑ Applicability to different stages of the policy cycle
- ☑ A transparent evaluation of scientific evidence – with the extent to which the likelihood of selection biases is reduced dependent on *how* certain steps are undertaken.

RECOMMENDATION 4. KEY FEATURES OF THE FRAMEWORKS

1) The CEE (Collaboration for Environmental Evidence) framework is the most detailed and represents a "gold standard". While it may not be possible for the Ministry to follow this framework to the letter, we recommend that Ministry staff familiarise themselves with the guidance provided by CEE and with some of the other resources referenced therein.
2) A good understanding of the CEE framework and guidance will provide Ministry staff with a good understanding of what an ideal review looks like and where / how adjustments and compromises can be made to speed up the process.
3) The CE (Conservation Evidence) framework is an interesting combination of systematic mapping and expert assessment of the evidence, which could be adapted to the needs of the Ministry by modifying the expert panel scoring system to answer a broader range of questions (it is only conceived to assess the effectiveness of conservation actions). However, it is important to keep in mind that, with this framework, the ultimate outcome of the synthesis (i.e., judgements on the effectiveness of conservation actions) is based entirely on expert opinion as there is no appraisal of individual evidence sources (as opposed to all other frameworks examined).
4) The EcoEvidence and US EPA frameworks can initially appear more user-friendly because they use straightforward appraisal systems and provide only high-level guidance for some of the evidence synthesis steps. However, it is important to consider that simple appraisal approaches may be too simplistic and may not provide the reviewer with enough guidance. We also recommend referring to the CEE guidelines when other frameworks do not provide sufficient detail and guidance.

This recommendation supports:

- ☑ Providing for an 'absolute' assessment of evidence
- ☑ Transferability of a process across policy questions in different environmental domains and at different levels of government – there remains flexibility as to *how* each of the steps is delivered to adapt to the different circumstances.
- ☑ Implementation of a repeatable process that allows for follow up assessments in the future.
- ☑ Applicability to different stages of the policy cycle
- ☑ A transparent evaluation of scientific evidence – with the extent to which the likelihood of selection biases is reduced dependent on *how* certain steps are undertaken.

RECOMMENDATION 5. SELECTING AN APPRAISAL TOOL FOR INDIVIDUAL STUDIES

1) Consider whether the use of numerical scores in certain tools is appropriate or not. Many authors and guidelines warn against doing so.
2) Beware of tools of unclear origin and do not assume that they have been correctly developed and tested.
3) Before developing a new tool or using one of unclear origin, consult with experienced topic experts and ensure that the tool satisfies the FEAT principles (FOCUSED, EXTENSIVE, APPLIED, TRANSPARENT).
4) Do not use hierarchies of evidence only as appraisal tools.

This recommendation supports:

- ☑ Providing for an 'absolute' assessment of evidence
- ☑ Transferability of a process across policy questions in different environmental domains and at different levels of government – there remains flexibility as to *how* each of the steps is delivered to adapt to the different circumstances.
- ☑ Implementation of a repeatable process that allows for follow up assessments in the future.
- ☑ Applicability to different stages of the policy cycle
- ☑ A transparent evaluation of scientific evidence – with the extent to which the likelihood of selection biases is reduced dependent on *how* certain steps are undertaken.

RECOMMENDATION 8. SELECTING AN APPRAISAL TOOL FOR REVIEWS

1) There are not many tools developed specifically to appraise environmental evidence syntheses. CEESAT appears to be the most comprehensive and reliable tool for this purpose.
2) Tools developed in other fields, like AMSTAR-2, may be applicable to the appraisal of environmental evidence syntheses.

This recommendation supports:

- ☑ Providing for an 'absolute' assessment of evidence
- ☑ Transferability of a process across policy questions in different environmental domains and at different levels of government – there remains flexibility as to *how* each of the steps is delivered to adapt to the different circumstances.
- ☑ Implementation of a repeatable process that allows for follow up assessments in the future.
- ☑ Applicability to different stages of the policy cycle
- ☑ A transparent evaluation of scientific evidence – with the extent to which the likelihood of selection biases is reduced dependent on *how* certain steps are undertaken.

RECOMMENDATION 7. COMMUNICATING THE FINDINGS

It is recommended that the following aspects are considered in the overall design, implementation, and resourcing of a potential process for assessing the strength of scientific evidence and communicating those findings to policymakers.

1) That the findings of an evidence synthesis are structured in a way that enables a policymaker to quickly identify the findings of the synthesis (e.g., in a short executive summary), while also providing further information and links for more of a 'deep dive' on particular aspects of the synthesis.
2) That the findings of an evidence synthesis include (both in the summary and in the main body of the findings):
    a) The policy problem / question; interventions assessed, and outcomes measured.
    b) The evidence synthesis method followed.
    c) A description of the method (and any limitations arising as a result of time/resource constraints).
    d) Types of evidence found.
    e) Variabilities in the evidence.
    f) Level of scientific rigour in the evidence.
    g) Limitations of the evidence.
    h) Effectiveness of the intervention at achieving the stated outcomes.
    i) Policy implications.
    j) Research gaps and recommendations for further primary research.
    k) References of interest.
3) That the method of communicating the findings of evidence synthesis ideally:
    a) Is multi-faceted (e.g., a combination or selection of documents, presentations and interactive sessions).
    b) Is supported with graphics and images.
    c) Uses lay language that is traceable to detailed scientific descriptions and original sources (e.g., by using hyperlinks).
    d) Is targeted to, and informed by an understanding of, the audience receiving the summaries.
4) That suitably trained science communicators with the appropriate specialist skills are involved as part of the team in delivering an evidence synthesis.
5) That summaries are provided in a form that meets the needs of practitioners – reinforcing the need for science communicators to be involved in planning stages of the synthesis project.

To assist with scenarios where there are significant time / resource constraints, there may be some benefit in sourcing some examples of high-quality summaries of evidence synthesis that could be translated into 'templates'.

These recommendations may assist with improving communication and dissemination skills between scientists and decision-makers.

4. References

Adams, W. M., & Sandbrook, C. (2013). Conservation, evidence and policy. *Oryx*, *47*(3), 329–335.

Bilotta, G. S., Milner, A. M., & Boyd, I. L. (2014). Quality assessment tools for evidence from environmental science. *Environmental Evidence*, *3*, 1–14.

Borenstein, M. (2009). Effect sizes for continuous data. In *The Handbook of Research Synthesis and Meta-Analysis* (2nd ed., pp. 221–235). Russell Sage Foundation.

Bowen, S., & Zwi, A. B. (2005). Pathways to "evidence-informed" policy and practice: A framework for action. *PLoS Medicine*, *2*(7), e166.

Brisco, E., Kulinskaya, E., & Koricheva, J. (2023). Assessment of temporal instability in the applied ecology and conservation evidence base. *Research Synthesis Methods*, 1–15.

Christie, A. P., Downey, H., Frick, W. F., Grainger, M., O'Brien, D., Tinsley-Marshall, P., White, T. B., Winter, M., & Sutherland, W. J. (2022). A practical conservation tool to combine diverse types of evidence for transparent evidence-based decision-making. *Conservation Science and Practice*, *4*(1), e579.

Christie, A. P., Morgan, W. H., Salafsky, N., White, T. B., Irvine, R., Boenisch, N., Chiaravalloti, R. M., Kincaid, K., Rezaie, A. M., & Yamashita, H. (2023a). Assessing diverse evidence to improve conservation decision-making. *Conservation Science and Practice*, *5*(10), e13024.

Christie, A. P., Morgan, W. H., Salafsky, N., White, T. B., Irvine, R., Boenisch, N., Chiaravalloti, R. M., Kincaid, K., Rezaie, A. M., & Yamashita, H. (2023b). Assessing diverse evidence to improve conservation decision-making. *Conservation Science and Practice*, *5*(10), e13024.

Cook, C. N., Nichols, S. J., Webb, J. A., Fuller, R. A., & Richards, R. M. (2017). Simplifying the selection of evidence synthesis methods to inform environmental decisions: A guide for decision makers and scientists. *Biological Conservation*, *213*, 135–145.

Cooke, S. J., Cook, C. N., Nguyen, V. M., Walsh, J. C., Young, N., Cvitanovic, C., Grainger, M. J., Randall, N. P., Muir, M., & Kadykalo, A. N. (2023). Environmental evidence in action: On the science and practice of evidence synthesis and evidence-based decision-making. *Environmental Evidence*, *12*(1), 10.

Dicks, L. V., Hodge, I., Randall, N. P., Scharlemann, J. P. W., Siriwardena, G. M., Smith, H. G., Smith, R. K., & Sutherland, W. J. (2014). A transparent process for "evidence-informed" policy making. *Conservation Letters*, *7*(2), 119–125.

Dicks, L. V., Walsh, J. C., & Sutherland, W. J. (2014). Organising evidence for environmental management decisions: A '4S'hierarchy. *Trends in Ecology & Evolution*, *29*(11), 607–613.

Environmental Protection Agency. (2016). *Weight of evidence in ecological assessment*. Environmental Protection Agency.

Frampton, G., Whaley, P., Bennett, M., Bilotta, G., Dorne, J.-L. C. M., Eales, J., James, K., Kohl, C., Land, M., & Livoreil, B. (2022). Principles and framework for assessing the risk of bias

for studies included in comparative quantitative environmental systematic reviews. *Environmental Evidence*, *11*(1), 12.

Gates, S. (2002). Review of methodology of quantitative reviews using meta-analysis in ecology. *Journal of Animal Ecology*, *71*(4), 547–557.

Gilbert, R., Salanti, G., Harden, M., & See, S. (2005). Infant sleeping position and the sudden infant death syndrome: Systematic review of observational studies and historical review of recommendations from 1940 to 2002. *International Journal of Epidemiology*, *34*(4), 874–887.

Gurevitch, J., & Hedges, L. V. (2020). Meta-analysis: Combining the results of independent experiments. In *Design and Analysis of Ecological Experiments* (pp. 378–398). Chapman and Hall/CRC.

Haddaway, N. R., & Westgate, M. J. (2019). Predicting the time needed for environmental systematic reviews and systematic maps. *Conservation Biology*, *33*(2), 434–443.

Haddaway, N. R., Woodcock, P., Macura, B., & Collins, A. (2015). Making literature reviews more reliable through application of lessons from systematic reviews. *Conservation Biology*, *29*(6), 1596–1605.

Haug, C., Rayner, T., Jordan, A., Hildingsson, R., Stripple, J., Monni, S., Huitema, D., Massey, E., van Asselt, H., & Berkhout, F. (2010). Navigating the dilemmas of climate policy in Europe: Evidence from policy evaluation studies. *Climatic Change*, *101*, 427–445.

Hendriks, F., Kienhues, D., & Bromme, R. (2016). Trust in science and the science of trust. In *Trust and Communication in a Digitized World, Models and Concepts of Trust Research*. Springer.

Hill, D., & Arnold, R. (2012). Building the evidence base for ecological impact assessment and mitigation. *Journal of Applied Ecology*, *49*(1), 6–9.

HM Treasury. (2020). *Magenta book Annex A: Analytical methods for use within an evaluation*. HM Treasury. https://assets.publishing.service.gov.uk/media/5e96c41a86650c2dd9e792ea/Magenta_Book_Annex_A._Analytical_methods_for_use_within_an_evaluation.pdf

Hosking, G. (2019). The decline of trust in government. In *Trust in Contemporary Society*. Brill.

Hunter, S. B., zu Ermgassen, S. O. S. E., Downey, H., Griffiths, R. A., & Howe, C. (2021). Evidence shortfalls in the recommendations and guidance underpinning ecological mitigation for infrastructure developments. *Ecological Solutions and Evidence*, *2*(3), e12089.

INASP. (2016). *Evidence informed policy making toolkit*. INASP. https://www.inasp.info/sites/default/files/2018-04/EIPM%20Toolkit-Ed2-FULL.pdf

Kadykalo, A. N., Cooke, S. J., & Young, N. (2021). The role of western-based scientific, indigenous and local knowledge in wildlife management and conservation. *People and Nature*, *3*(3), 610–626.

Kano, H., & Hayashi, T. I. (2021). A framework for implementing evidence in policymaking: Perspectives and phases of evidence evaluation in the science-policy interaction. *Environmental Science & Policy*, *116*, 86–95.

Kenny, C., Hobbs, A., Tyler, C., & Blackstock, J. (2018). *POST Research Study: The work and impact of POST*. STEaPP and Economic & Social Research Council.

Koolen-Bourke, D., & Peart, R. (2022). *Science for policy: The role of science in the national policy statement for freshwater management*. Environmental Defence Society. https://eds.org.nz/wp-content/uploads/2022/08/Freshwater-Policy-Report_FINAL_CorrectedPostLaw-Suit.pdf

Koricheva, J., Gurevitch, J., & Mengersen, K. (2013). *Handbook of meta-analysis in ecology and evolution*. Princeton University Press.

Koricheva, J., Jennions, M., & Lau, J. (2013). Temporal trends in effect sizes: Causes, detection, and implications. In *Handbook of Meta-analysis in Ecology and Evolution*. Princeton University Press.

Macura, B., Suškevičs, M., Garside, R., Hannes, K., Rees, R., & Rodela, R. (2019). Systematic reviews of qualitative evidence for environmental policy and management: An overview of different methodological options. *Environmental Evidence*, *8*, 1–11.

Miljand, M., & Eckerberg, K. (2022). Using systematic reviews to inform environmental policy-making. *Evaluation*, *28*(2), 210–230.

Mupepele, A., Walsh, J. C., Sutherland, W. J., & Dormann, C. F. (2016). An evidence assessment tool for ecosystem services and conservation studies. *Ecological Applications*, *26*(5), 1295–1301.

Nakagawa, S., Yang, Y., Macartney, E. L., Spake, R., & Lagisz, M. (2023). Quantitative evidence synthesis: A practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environmental Evidence*, *12*(1), 8.

Nichols, S. J., Peat, M., & Webb, J. A. (2017). Challenges for evidence-based environmental management: What is acceptable and sufficient evidence of causation? *Freshwater Science*, *36*(1), 240–249.

Nichols, S., Webb, A., Norris, R., & Stewardson, M. (2011). *Eco Evidence analysis methods manual: A systematic approach to evaluate causality in environmental science*. eWaterCRC. https://toolkit.ewater.org.au/Tools/Eco-Evidence/documentation

Norris, S. L., Aung, M. T., Chartres, N., & Woodruff, T. J. (2021). Evidence-to-decision frameworks: A review and analysis to inform decision-making for environmental health interventions. *Environmental Health*, *20*, 1–33.

OECD. (2020). Standards of evidence: Mapping the experience in OECD countries. In *Mobilising Evidence for Good Governance: Taking Stock of Principles and Standards for Policy Design, Implementation and Evaluation*. OECD. https://www.oecd-ilibrary.org/sites/e361aae5-en/index.html?itemId=/content/component/e361aae5-en

Office of the Ombudsman. (2019). *The OIA and the public policy making process*. Office of the Ombudsman. https://www.ombudsman.parliament.nz/sites/default/files/2020-05/The%20OIA%20and%20the%20public%20policy%20making%20process%20August%202019%20.pdf

O'Leary, B. C., Kvist, K., Bayliss, H. R., Derroire, G., Healey, J. R., Hughes, K., Kleinschroth, F., Sciberras, M., Woodcock, P., & Pullin, A. S. (2016). The reliability of evidence review methodology in environmental science and conservation. *Environmental Science & Policy*, *64*, 75–82.

Pe'er, G., Bonn, A., Bruelheide, H., Dieker, P., Eisenhauer, N., Feindt, P. H., Hagedorn, G., Hansjürgens, B., Herzon, I., & Lomba, Â. (2020). Action needed for the EU Common Agricultural Policy to address sustainability challenges. *People and Nature*, *2*(2), 305–316.

Perry, R., Whitmarsh, A., Leach, V., & Davies, P. (2021). A comparison of two assessment tools used in overviews of systematic reviews: ROBIS versus AMSTAR-2. *Systematic Reviews*, *10*, 1–20.

Pullin, A. S., Cheng, S. H., Jackson, J. D., Eales, J., Envall, I., Fada, S. J., Frampton, G. K., Harper, M., Kadykalo, A. N., & Kohl, C. (2022). Standards of conduct and reporting in evidence syntheses that could inform environmental policy and management decisions. *Environmental Evidence*, *11*(1), 16.

Pullin, A. S., & Knight, T. M. (2003). Support for decision making in conservation practice: An evidence-based approach. *Journal for Nature Conservation*, *11*(2), 83–90.

Pullin, A. S., & Knight, T. M. (2005). Assessing conservation management's evidence base: A survey of management-plan compilers in the United Kingdom and Australia. *Conservation Biology*, *19*(6), 1989–1996.

Randall, N., & Donnison, L. (2014). *The Value of On-farm Interventions for Improving Water Quality. What is the Evidence?* Department for Environment, Food and Rural Affairs. https://www.researchgate.net/publication/260408118_The_Value_of_On-farm_Interventions_for_Improving_Water_Quality_What_is_the_Evidence/link/004635310a40744f20000000/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19

Randall, N., Donnison, L., Lewis, P., & James, K. L. (2015). *How effective are on-farm mitigation measures for delivering an improved water environment? A systematic map*.

Reed, M., & Meagher, L. (2019). Using evidence in environmental and sustainability issues. In *What works now? Evidence-informed policy and practice*. Policy Press.

Rowland, J. A., Bracey, C., Moore, J. L., Cook, C. N., Bragge, P., & Walsh, J. C. (2021). Effectiveness of conservation interventions globally for degraded peatlands in cool-climate regions. *Biological Conservation*, *263*, 109327.

Salafsky, N., Boshoven, J., Burivalova, Z., Dubois, N. S., Gomez, A., Johnson, A., Lee, A., Margoluis, R., Morrison, J., & Muir, M. (2019). Defining and using evidence in conservation practice. *Conservation Science and Practice*, *1*(5), e27.

Salafsky, N., Irvine, R., Boshoven, J., Lucas, J., Prior, K., Bisaillon, J., Graham, B., Harper, P., Laurin, A. Y., & Lavers, A. (2022). A practical approach to assessing existing evidence for specific conservation strategies. *Conservation Science and Practice*, *4*(4), e12654.

Sarkki, S., Niemelä, J., Tinch, R., Van Den Hove, S., Watt, A., & Young, J. (2014). Balancing credibility, relevance and legitimacy: A critical assessment of trade-offs in science–policy interfaces. *Science and Public Policy*, *41*(2), 194–206.

Schwartz, M. W., Cook, C. N., Pressey, R. L., Pullin, A. S., Runge, M. C., Salafsky, N., Sutherland, W. J., & Williamson, M. A. (2018). Decision support frameworks and tools for conservation. *Conservation Letters*, *11*(2), e12385.

Singh, G. G., Lerner, J., Mach, M., Murray, C. C., Ranieri, B., St-Laurent, G. P., Wong, J., Guimaraes, A., Yunda-Guarin, G., Satterfield, T., & Chan, K. M. A. (2020). Scientific shortcomings in environmental impact statements internationally. *People and Nature*, *2*(2), 369–379.

Stanhope, J., & Weinstein, P. (2023). Critical appraisal in ecology: What tools are available, and what is being used in systematic reviews? *Research Synthesis Methods*, *14*(3), 342–356.

Superu. (2018). *Making sense of evidence: A guide to using evidence in policy: Using evidence for impact*. Social Policy Evaluation and Research Unit (Superu). https://thehub.swa.govt.nz/assets/Uploads/Making-Sense-of-Evidence-handbook-FINAL.pdf

Suter, G., Cormier, S., & Barron, M. (2017). A weight of evidence framework for environmental assessments: Inferring qualities. *Integrated Environmental Assessment and Management*, *13*(6), 1038–1044.

Sutherland, W. J. (Ed.). (2022). *Transforming conservation: A practical guide to evidence and decision making*. Open Book Publishers. https://www.openbookpublishers.com/books/10.11647/obp.0321

Sutherland, W. J., Dicks, L. V., Petrovan, S. O., & Smith, R. K. (Eds.). (2020). *What works in conservation 2020*. Open Book Publishers. https://books.openbookpublishers.com/10.11647/obp.0191.pdf

Sutherland, W. J., Downey, H., Frick, W. F., Tinsley-Marshall, P., & McPherson, T. (2021). Planning practical evidence-based decision making in conservation within time constraints: The Strategic Evidence Assessment Framework. *Journal for Nature Conservation*, *60*, 125975.

Sutherland, W. J., & Wordley, C. F. R. (2018). A fresh approach to evidence synthesis. *Nature*, *558*, 364–366.

Sutton, A., Clowes, M., Preston, L., & Booth, A. (2019). Meeting the review family: Exploring review types and associated information retrieval requirements. *Health Information & Libraries Journal*, *36*(3), 202–222.

United Nations Environment Programme. (2020). *Assessment of options for strengthening the science-policy interface at the international level for the sound management of chemicals and waste*. United Nations Environment Programme. https://wedocs.unep.org/bitstream/handle/20.500.11822/33808/OSSP.pdf

Walsh, J. C., Dicks, L. V., Raymond, C. M., & Sutherland, W. J. (2019). A typology of barriers and enablers of scientific evidence use in conservation practice. *Journal of Environmental Management*, *250*, 109481.

Walsh, J. C., Dicks, L. V., & Sutherland, W. J. (2015). The effect of scientific evidence on conservation practitioners' management decisions. *Conservation Biology*, *29*(1), 88–98.

Warner, C. (2022, April 11). *Perspectives in public policy—The policy cycle*. The University of Auckland. https://www.online.auckland.ac.nz/2022/04/11/perspectives-in-public-policy-the-policy-cycle/

Wood, C. (n.d.). *Communicating evidence to policy makers—What works best?* Chartered Institute of Public Relations.

Woodcock, P., Pullin, A. S., & Kaiser, M. J. (2014). Evaluating and improving the reliability of evidence syntheses in conservation and environmental science: A methodology. *Biological Conservation*, *176*, 54–62.

Appendix 1: Outline of the CEE method[1]

[1] Guidelines for Authors – Environmental Evidence

1. Planning a synthesis

The rigour and transparency of an evidence synthesis starts in its planning phase. At this time, the involvement of relevant stakeholders in Steps 1a-e described below is essential to inform the goals and structure of the review and to ensure that the process of evidence synthesis is as free from bias as possible. For complex reviews, advisory groups or panels may be consulted during the planning stage.

a) **Defining the question to be answered.** Questions appropriate for systematic reviews need to be specific, well defined, and relatively simple (i.e., closed-framed questions that require answers from a set of predefined responses), while systematic maps are better suited to broader (open-framed) questions (Table 1).

Key-elements making a question suitable for evidence synthesis are currently referred to using the Population, Intervention, Comparator and Outcomes (PICO) or Population, Exposure, Comparator and Outcomes (PECO) acronyms (Table 2).

In the example in Table 2, the PECO elements are:

- Population = endemic European birds
- Exposure = motorways within habitat
- Comparator = habitats without motorways
- Outcomes = breeding success

Table 1: Example of question formulation provided by Collaboration for Environmental Evidence[2]

Question	Key elements	Question type
Starting question: What is the impact of roads on wildlife?	None specified other than roads (vague), wildlife (vague) and impact (vague)	Open-framed (possible for Systematic Mapping but unsuitable for Systematic Review)
Refined question: What is the impact of motorways on populations of endemic bird species in Europe?	Motorways (=exposure), European endemic bird species (=population); but comparator and outcome not specified	Open-framed (suitable for Systematic Mapping but unsuitable for Systematic Review)
Further refined question: What is the impact of habitats containing motorways on the breeding success of endemic European bird species, as compared to habitats without motorways?	Motorways (=exposure), no motorways (=comparator), European endemic bird species (=population), breeding success (=outcome)	Closed-framed (suitable for Systematic Mapping and possible for Systematic Review)

[2] https://environmentalevidence.org/information-for-authors/2-need-for-evidence-synthesis-type-and-review-team-2/

Table 2: Elements of a PICO / PECO question[3]

Question element	Definition
Population (of subjects)	Unit of study (e.g. ecosystem, species) that should be defined in terms of the statistical populations of subject(s) to which the intervention will be applied.
Intervention/exposure	Proposed management regime, policy, action or the environmental variable to which the subject populations are exposed.
Comparator	Either a control with no intervention/exposure or an alternative intervention or a counterfactual scenario.
Outcome	All relevant outcomes from the proposed intervention or environmental exposure that can be reliably measured

b) **Scoping the evidence.** Before any further steps, it is essential that a preliminary scoping of the evidence is undertaken to guide the selection of the synthesis method and the development of the review protocol. A scoping exercise can provide a first indication of the amount and type of evidence available and of the likely extent and reliability of the findings. This information can be used to inform the following steps of the process and, if needed, to reconsider the original question and objectives of the project.

c) **Estimating resource requirements and timelines.** Scoping should provide an estimate of the time effort required for the review, so that a realistic budget can be prepared. Systematic / rapid reviews and systematic maps are inevitably time-consuming, but significant time savings can be achieved with good planning, adequate resource, and a skilled review team.

d) **Choosing the synthesis method.** As the question is formulated, it should become clear whether the question can be answered using a systematic review (which can be shortened into a rapid review) or a systematic map. Systematic reviews and systematic maps share the same initial steps but differ in their analytical approaches and outputs (Table 3). Systematic maps are often preliminary syntheses of the evidence relating to a broader question. If sufficient evidence is available for further synthesis, the question may then be refined and made more specific to inform a systematic review (or an abbreviated rapid form).

[3] https://environmentalevidence.org/information-for-authors/2-need-for-evidence-synthesis-type-and-review-team-2/

Table 3. Key similarities and differences between systematic reviews and systematic maps[4].

	Systematic Review	Systematic Map
Protocol	Mandatory	Mandatory
Systematic searching	Mandatory	Mandatory
Systematic study selection	Mandatory	Mandatory
Critical appraisal of study validity	Mandatory, to ensure robustness of the review answer – directly influences the data synthesis and interpretation steps	Optional (possible if study validity indicators can be captured using the coding method) – does not influence mapping process itself
Data coding and extraction	Mandatory, Meta-data coded and outcome measures (e.g. effect sizes) extracted.	Mandatory, metadata only coded. No extraction of outcome measures (e.g. effect sizes).
Data synthesis approach	Aggregative, seeking an unbiased answer with known precision; could involve meta-analysis	Exploratory; may include coding and group analysis
Typical output	A quantitative or qualitative answer with an indication of uncertainty and any threats to validity. May include estimate of variance caused by external factors.	A description of the evidence base, showing the distribution and abundance of evidence across different elements of the question. A relational database may be provided.

e) **Assembling the review team.** Systematic / rapid reviews and systematic maps are time-consuming and usually require a multidisciplinary team. The team may include subject experts working alongside review and synthesis methodology experts. A team of reviewers is needed not only to provide all the required technical expertise, but also because several stages of the review process must be undertaken by at least two people to minimise the risk of introducing errors or bias.

The review team should be led by a Lead Reviewer who is experienced in the review methodology. The inclusion of subject experts in the team is needed but brings with it the potential for bias. Careful consideration should be given to the independence of subject experts and conflicts of interest should be declared and avoided where possible.

To further ensure independence of conduct and avoid conflicts of interest, the members of the review team must not be the commissioner of the work, the ultimate users of the evidence, or stakeholders in the process.

2. Developing a protocol

The protocol for the evidence synthesis is an independent document to be prepared before the synthesis is conducted. The protocol serves as a guide and reference to the conduct of the synthesis, which should reflect the views of all the parties involved in the planning phase (i.e.,

[4] https://environmentalevidence.org/information-for-authors/2-need-for-evidence-synthesis-type-and-review-team-2/

the commissioner of the work, the ultimate users of the evidence, the stakeholders in the process and the review team).

The protocol is essential to minimise reviewer bias. Any diversion from the protocol during the synthesis process is discouraged. However, when changes to the original methodology are necessary, these must be compulsorily recorded and motivated. This is particularly important to maintain transparency and repeatability, as well as the confidence of users of the evidence and stakeholders.

The structure of an evidence synthesis protocol mirrors the structure of the systematic / rapid review or systematic map that it guides. The protocol should outline:

- The problem being addressed and the need for an evidence synthesis. It is important to clearly outline the mechanisms whereby the intervention or activity is thought to have an impact on a specific population / natural system.

- The strategy for searching for relevant studies. This should include a discussion of the criteria defining when to stop the search if resource constraints (such as time, people power, skills) are anticipated.

- The eligibility criteria for screening the studies. These criteria are in part defined by the way in which the question to be answered was formulated and in part by decisions about the kinds of evidence (e.g., study designs) deemed suitable for inclusion in the synthesis.

- The methods to be used for data coding / extraction, study validity assessment, and data synthesis.

- Conflicts of interest and funding sources.

CCE provides protocol templates[5,6] as well as the option to register an evidence synthesis protocol in PROCEED, an open access registry of titles and protocols for prospective evidence syntheses in the environmental sector. This is not mandatory but is considered important to avoid duplication of effort and to reduce risk of bias in the conduct of reviews by encouraging the practice of protocol development[7]. The registration of a protocol in PROCEED is free and includes feedback from the editors of the portal.

3. Conducting a search

Searches should be transparent and reproducible. In practice, it is unlikely that absolutely all the relevant literature can be identified during the search, but a key requirement is to try to gather as much of the available evidence as possible to minimise bias in the findings. Any limitations of the search, such as lack of access to or inability to use some literature (for example because of a language barrier) should be clearly reported. Enlisting an information specialist in the review team is recommended to establish an efficient search strategy. A good search strategy can also make a substantial difference to the time and cost of a synthesis. In addition, because of the

[5] https://environmentalevidencejournal.biomedcentral.com/submission-guidelines/preparing-your-manuscript/systematic-review-protocol

[6] https://environmentalevidencejournal.biomedcentral.com/submission-guidelines/preparing-your-manuscript/systematic-map-protocol

[7] https://environmentalevidence.org/proceed/

systematic aspect of the searching and the need to keep careful track of the findings, review teams should, when possible, include librarians or information specialists.

Key aspects of the search process are outlined below, and more information is available as part of the CCE guidelines[8].

a) **Avoiding search errors.** Errors that can occur during the search include missing search terms, unintentional misspelling of search terms, errors in the search syntax when combining multiple search terms into search strings, and use of inappropriate search terms. These errors can be minimised with a well-developed protocol and internal peer-review within the review team.

b) **Avoiding search biases.** Some of the most common sources of systematic bias include:

- Language bias: studies written in English are more likely to be searched and accessed. When possible, it is recommended to look beyond the English language literature.

- Prevailing paradigm bias: studies relating to or supporting the prevailing paradigm or topic (for example climate change) are more likely to be published and hence discoverable. To maintain the search specificity, it is important to ensure that studies properly meet the search criteria without simply referring to the topic of interest.

- Temporal bias: older articles may be overlooked in favour of more recent evidence. Therefore, it is important to also search older publications. In addition, studies supporting a novel development (for example a new hypothesis or methodology) are more likely to be published soon after the novel development, but their results may not be confirmed by subsequent studies. Therefore, it is important to consider updating the search in the future.

- Publication bias: statistically significant results (positive results) are more likely to be accepted for publication than non-significant ones (negative results). This may lead to overestimating the effect/impact of the factor/variables being tested. To minimise this bias, searches for studies reporting non-significant results should be conducted in the grey literature (theses, conference papers and reports) and unpublished datasets should be examined.

For a more extensive analysis of search biases see (Bayliss & Beyer, 2015).

c) **Establishing a test-list.** Before, the search, the review team should collate a set of resources relevant to answer the question of the evidence synthesis. The test-list is used to assess the performance of the search strategy.

d) **Identifying search terms and developing search strings.** Initial search terms can usually be generated from the question elements and by looking at the resources in the test-list. However, the full range of the PICO / PECO criteria may not always be identifiable in the title and abstract of a study paper. As a consequence, building search strings (combinations of key words and phrases) from search terms requires project teams to draw upon both their scientific expertise, a certain degree of imagination, and an analysis of titles and abstracts to consider how authors might use different terminologies to describe their research. This is an

[8] https://environmentalevidence.org/information-for-authors/4-conducting-a-search/

iterative process, testing search strings using selected databases, recording numbers of references identified, and sampling titles for proportional relevance or specificity.

e) **Searching different types of sources.** Several sources should be searched to ensure that as many relevant articles as possible are identified. Sources need to be selected based on the disciplines addressed by the question driving the synthesis. The capacity of sources to provide the greatest quantity of relevant articles for a limited number of searches and their susceptibility to search biases also needs to be considered.

   An approach commonly recommended is to start the search using the source where the largest number of relevant papers are likely to be found, and subsequent searches can be constructed with the aim to complement these first results. Sources containing abstracts allow greater understanding of relevance and should be given priority.

   Examples of different types of sources (of mostly academic research) include:

   - Web of Science
   - Scopus
   - Google Scholar
   - ResearchGate
   - BASE Bielefeld academic search engine
   - Publishers' websites (e.g., Elsevier's ScienceDirect and Wiley Interscience).

   Examples of sources of grey literature include:

   - www.greynet.org
   - University libraries
   - Websites of organisations and professional networks
   - Search engines (Google)
   - Consultation of technical experts

f) **Stopping the search.** The criteria to stop the search should be pre-defined and outlined in the review protocol. While time and budget are often major constraints, ideally the right moment for stopping the search is only when additional unit of time spent searching returns progressively fewer relevant references. If relevant resources are being identified, the search should continue. Statistical techniques, such as capture-recapture and the relative recall method, exist to guide decisions about when to stop searching, although these to not appear to have been widely used.

g) **Keeping track of the search details.** The search methodology should be thoroughly documented *a-priori* in the review protocol. During the search, enough detail should be recorded to allow the search to be replicated including the name of the sources searched,

the date of the search and the search terms / strings used. The search history and number of articles retrieved by each search should be recorded in a logbook or using screenshots.

h) **Reporting.** All relevant information about the search, including the results and performances of the search and any amendment made to the original protocol, should be reported in the final evidence synthesis report, possibly as additional files, or supplementary information. The limitations of the search should be clearly outlined, including the range of languages, types of documents, time-period covered by the search, date of the search and any unexpected difficulty that impacted the search compared to what was described in the original protocol.

i) **Updating or amending the search.** Searches may need to be updated or amended as new evidence becomes available. Thorough documentation of the search protocols will allow this work to be undertaken by a different review team if needed.

Measures to speed up the search for rapid reviews (all should be documented and justified):

- Including date, language, geographical limitations.
- Searching only key databases.

4. Screening the evidence

Eligibility criteria are used as part of a systematic screening process to establish whether the resources identified by the search are relevant for answering the question driving the evidence synthesis. Both the eligibility criteria and the screening process should be planned in advance and specified in the evidence synthesis protocol.

A typical approach to ensure consistency within the review team is to develop an eligibility screening form containing the eligibility criteria along with instructions for the reviewers so that each reviewer follows the same procedure. The eligibility criteria and the screening process should be pilot-tested and refined as part of the development of the protocol for the evidence synthesis.

a) **Eligibility criteria.** The use of pre-specified and explicit eligibility criteria ensures that the inclusion or exclusion of the resources identified by the search is done in an objective and transparent manner. The eligibility criteria should be few and easy to locate within the resources being screened.

   The eligibility criteria should reflect the question that the review is trying to answer and, therefore, follow logically from the PICO / PECO elements that define the question structure. The PICO / PECO elements (Population, Impact / Effect, Comparator, Outcomes; refer to **Error! Reference source not found.**2) must be clearly identifiable for a study to be eligible for inclusion in the evidence synthesis. In addition to the question to be answered, other phases of a synthesis contributing to shaping the eligibility criteria are the initial scoping of the evidence (see Step 1) and the development of the search strategy (see Step 2).

   Finally, the study design (e.g., observational, or experimental) should be included among the eligibility criteria. The design of the studies retained during the screening process should be compatible with the planned approach for the data synthesis of the review (see Step 7). Some study designs may also be more prone to bias than others, but a full assessment of

the risks of bias and other threats to validity takes place after the screening process, at the critical appraisal step (see Step 69).

b) **The screening process.** The screening process ensures that the eligibility criteria are applied consistently. The process normally involves two steps:

- A first screening based on titles and abstracts to remove articles which are clearly irrelevant.
- A further assessment of the full text of the resource.

Resources based on the same study (i.e., linked articles) should be grouped together and screened for eligibility as a single unit, unless they fully overlap (in which case duplicates should be removed).

Records of all screening decisions should be kept (in a database or reference management tool), so that the judgements made are transparent and defensible.

c) **The screening team.** The screening process involves judgement and should be conducted carefully. The screening should be performed where possible by at least two people. The screeners do not necessarily need to be the same for all the resources or for all screening steps. Ideally, both screeners should independently perform the selection process and then compare their decisions. Alternatively, one person can be in charge of the selection process and a second person act as a reviewer by checking the screening decisions. In this case, the reviewer must examine an adequate number of resources / decisions.

The use of a single screener is not considered best practice. However, if proceeding with a single screener becomes necessary, this should be recorded in the synthesis protocol and final report along with a discussion of the reasons and implications of the deviation from the standard two-screener system.

An assessment of agreement between screeners helps to ensure that the screening process is reproducible and reliable, therefore, it is important to record any disagreement. A process for resolving disagreements should be included in the synthesis protocol. A simple approach involves discussions between the screeners to reach a consensus. A third opinion from another member of the review team or from the project advisory group can be sought if needed.

Measures to speed up the screening for rapid reviews (all should be documented and justified):

- Using only one screener, but as many references as possible should be dual-screened and the consistency of screening decisions should be tested.
- Using eligibility criteria that place emphasis on higher validity study designs.

5. Data coding and extraction

Data coding and extraction refer to the process of systematically extracting relevant information from the resources retained following the screening process.

- Data coding is the recording of relevant characteristics (meta-data) of the study such as when and where the study was conducted and by whom, as well as aspects of the study design and conduct.

- Data extraction is only required for systematic reviews and refers to the recording of the results of the study (e.g., effect size, means and variances or other important findings). Data extraction.

To standardise and document the processes of data coding and extraction, a standard data coding or extraction form or table (e.g., spreadsheet) is usually developed, and tested, as part of the preparation of the protocol for the evidence synthesis. The spreadsheet contains prompts to help the reviewers to record all relevant information necessary to address the synthesis question, plus any additional information required for the critical appraisal of the resources (see Step 6). The final data coding or extraction table should be included in the evidence synthesis protocol. Data coding or extraction tables for systematic reviews are likely to be more detailed than for systematic maps.

As for the screening process, the data coding and extraction process should involve an element of peer-review, with one member of the review team undertaking coding and extraction and another person in the team checking at least a subset of the coded / extracted information.

Measures to speed up data coding and extraction for rapid reviews (all should be documented and justified):

- Using only one reviewer, but a sample of resources should be examined independently by two reviewers to test for consistency.

- Limiting coding and extraction to data necessary for the synthesis.

6. Critical appraisal of the evidence

In the critical appraisal stage, the resources retained following the screening process are assessed for their reliability for answering the question motivating the systematic review. Since the quality of scientific evidence varies considerably, the critical appraisal step is essential to identify the flaws and limitations of the evidence being used so that these can be considered when drawing the conclusions of the synthesis.

The appraisal of the evidence needs to consider two key elements:

- Internal validity. Internal validity refers to the extent of bias in the results of an individual study due to flaws in study design or conduct. The extent of bias can be inferred by examining the study design and methods to determine whether adequate steps were taken to protect against bias.

- External validity. Whilst internal validity is a specific property of an individual research study, external validity is context dependent. External validity is the extent to which the results of an individual study can be generalised and applied to other circumstances. This includes the suitability of the findings of a study for answering the question being addressed by the review. For example, how well do the results of control laboratory trials apply to answering a question related to effects / impacts occurring in the real world?

The critical appraisal process should be planned, and tested, while developing the protocol for the evidence synthesis. Key aspects of the search process are outlined in Step 6a-f, below and more information is available as part of the CCE guidelines[9].

a) **Preparing the team.** The team should be familiar with the strengths and weaknesses of research studies relevant to the review question and understand the concepts of internal and external validity. There should be enough people to allow dual assessments of each study and to ensure that reviewers are not required to assess studies of which they are authors or contributors.

b) **Identifying eligible study designs and possible sources of bias affecting internal validity.** Scoping searches conducted as part of the protocol development process (see Step 2) should reveal the types of studies that are likely to meet the review's eligibility criteria. The review team will need to be familiar with all eligible study designs to be able to decide which classes of bias and risk of bias tools may be relevant. Other potential sources of bias, in addition to design type, should also be identified and considered during the protocol development process.

c) **Selecting risk of bias tools.** Numerous checklists and risk of bias tools are available for evidence appraisal (see Appendix 4[10] in Frampton et al. (2022) for a sample list). Whilst many of these tools have been rigorously developed and tested, not all of them can be assumed to be fit-for-purpose (see more details in Frampton et al. (2022)). CEE has developed a prototype Critical Appraisal Tool for assessing bias in environmental research studies addressing PICO / PECO questions[11].This tool has not yet been widely tested and is subject to further revisions.

   When using any tool, the review team should carefully consider whether it fully captures all potential biases relevant to the resources included in the evidence synthesis. For evidence synthesis including both experimental and observational studies it may be necessary to use more than one critical appraisal tool. The review team should have sufficient expertise to identify the appropriate tools and to modify existing tools or develop new ones if needed. All these actions need to be defensible and carefully documented.

d) **Assessing interval validity (i.e., assessing the risk of bias).** A fit-for-purpose risk of bias tool should make bias identification relatively straightforward, but it is possible to proceed even in the absence of a tool by basing the appraisal on the following types of biases:

   - Bias due to confounding. This bias arises due to uncontrolled variables (confounders) that influence both the impact / exposure and the outcome.

   - Bias in selection of subjects / areas (selection bias). This bias can be caused by unconscious or intentional non-random selection of samples or data to support prior beliefs of the investigator(s).

   - Bias due to misclassification of the exposure (misclassified comparison bias; it applies to observational studies only). This bias arises from misclassification or mismeasurement of

[9] https://environmentalevidence.org/information-for-authors/7-critical-appraisal-of-study-validity/

[10] https://static-content.springer.com/esm/art%3A10.1186%2Fs13750-022-00264-0/MediaObjects/13750_2022_264_MOESM4_ESM.docx

[11] https://environmentalevidence.org/cee-critical-appraisal-tool

the impact / exposure and / or of the comparator, which leads to a misrepresentation of the association between the impact / exposure and the outcome.

- Bias due to deviation from the planned impact / exposure in experimental studies (performance bias; it applies to experimental studies only). This bias arises from the alteration of the planned impact / exposure or comparator after that start of the experiment.

- Bias due to missing data (attribution bias). This can be considered as a type of selection bias. As a results of this bias, data about subjects or areas that were initially included in the study are not available for inclusion in the analysis of the effect estimate.

- Bias in measurement of outcomes (detection bias). This bias is caused from non-random differences in measurements of outcomes. Systematic errors in measurements of outcomes may occur if outcome data are measured differently between the exposure and comparator groups.

- Bias in selection of the reported result (reporting bias). This bias is caused by selective reporting of study findings.

- Bias due to an inappropriate statistical analysis approach. Referred to as “risk of outcome assessment biases” in the CEE tool. This bias is caused by errors in statistical methods applied within the individual studies included in review.

- Other risks of bias. Any bias related to the study design of interest that is not covered above.

Many ‘risk of bias’ tools guide the reviewer by asking “signalling questions” about the study methods based on the common forms of bias listed above. A more detailed analysis of these forms of bias can be found in Appendix 5[10] in Frampton et al. (2022).

A score must be assigned to each form of bias and an overall score of internal validity for each resource being appraised needs to be determined. The scoring system may be provided by the risk of bias tool being used or may be established by the review team. Below are some common risk of bias scoring systems:

- High / Low / Unclear risk of bias. This scoring systems is straightforward, and the results are easy to tabulate or present graphically (e.g., using a “traffic light” red / amber / green approach). However, it may be tempting for the reviewers to be less decisive and assign most studies to the “unclear” category. See Higgins et al. (2011) for a detailed description of the criteria defining each category.

- Definitely Low / Probably Low / Probably High / Definitely High risk of bias. This approach avoids the use of the “unclear” category, requiring that instances of insufficient information are recorded within the “Probably High Risk” category. See National Toxicology Program (2015) for more details.

- Low / Moderate / Serious / Critical risk of bias / No information. Here categories of “Low” and “Moderate” risk of bias should be interpreted in relation to how well the study matches an ideal target study design. See Higgins et al. (2019) for more details.

To determine the overall score of internal validity of the individual resource being appraised, the score assigned to the individual forms of bias need to be combined. If a study is judged to have low risk for all relevant types of bias, then it can be safely determined that the study has a low risk of bias. However, if the study is deemed to have a high risk of bias for at least one type of bias, this is sufficient to rate the study as having a high risk of bias. More complex scenarios are described in Table 4.

Table 4. Examples of study-level risk of bias classifications (i.e., the overall score of internal validity assigned to the individual resources being appraised) from Frampton et al. (2022). The references included in the table are Higgins et al. (2019), Jüni et al. (2016) and Sterne et al. (2016).

Risk of bias classification for each class of bias	Possible study-level risk of bias classification for a given outcome of interest
All classes of bias are judged to have low risk of bias, definitely low risk of bias, or probably low risk of bias for the outcome of interest	Low risk
At least one class of bias is judged to have high risk of bias, definitely high risk of bias, or probably high risk of bias for the outcome of interest	High risk
Some classes of bias are judged to have low risk, others unclear risk, but no classes are judged to have high risk for the outcome of interest	Unclear risk (to avoid an unclear judgement for the overall study it is preferable for each domain to reach a probably low risk or probably high risk judgement instead of an "unclear" judgement, where possible)
No information available for any classes of bias for the outcome of interest	Unclear risk (or no information)
Classes of bias are judged as having combinations of "moderate", "serious" or "critical" risks of bias (or other terminology) for the outcome of interest	Summarising judgements other than high/low/unclear may not be intuitively straightforward. A clear rationale should be provided, based on logic (i.e. the criteria should not be arbitrary). See recent Cochrane tools [58, 59, 81] for examples

Numerical scores are sometimes employed for assessing the risk of bias. However, categorical judgements with explanations provide better information. Numeric scores are inadvisable for summarising risk of bias for several reasons:

- Numeric scores may imply that different types of bias have equal weight or can be quantified relative to each other.
- Numeric scores may imply that mathematical operations can be performed on categories. This can result in a misleading account of risk of bias.
- Numeric scores reported in one systematic review may not have the same meaning as the scores reported in another review.

In the review protocol, the review team should carefully document the sources of bias investigated and the scoring systems used for the appraisal process. This information should also be provided in the final evidence synthesis report. Each internal validity judgement should be accompanied by a concise written justification to reduce subjectivity of interpretation. If any changes to the methods are required during the appraisal process, these should be clearly documented in the final review report as deviations from the protocol. In such cases, the updated methods must be applied to all studies included in the review.

e) **Assessing external validity.** Guidelines for the assessment of external validity are not well established, therefore, there is more flexibility for review teams to determine their own approach if this is done in transparent and defensible manner.

A pragmatic way to assess external validity proposed by CEE is to consider systematically how well the key elements of the study are being appraised (i.e., the PICO / PECO elements and other relevant aspects of the study design), and how well they match those of the review question as shown in Table 5 5. This is not a prescriptive approach, but rather a template to help reviewers to identify limitations to the external validity of a study (indicated by “No”

answers in Table 5). These limitations can then be considered in detail and reported as an overall external validity score (High / Low).

Table 5. Template for assessing external validity of the resources being appraised[12].

PECO / PICO elements and other aspects of the setting	**Applicability** Would the studied comparison be feasible (applicable) in the setting of the review question? **Yes / No**	**Transportability** Are the study characteristics sufficiently similar to those of the review question setting? [a] **Yes / No**	Overall external validity Rate as "high" if all answers in each row are "Yes". Rate as "low" if there are any "No" answers
Population			
Exposure or intervention			
Comparator			
Outcome			
Spatial scale			
Temporal scale			
Mediator variables [b]			
Other aspects of study design (add as needed…)			

[a] If there are differences but these are appropriately adjusted for statistically answer "yes" and document the rationale for this judgement

[b] Variables that are on the causal pathway between the exposure / intervention and the outcome

Finally, the external validity score needs to be combined with the internal validity score to determine the overall validity of the resource being appraised. Studies with low internal validity (i.e., high risk of bias) will have low overall validity independently of their external validity score. Studies with high internal validity (i.e., low risk of bias) may still be assigned an overall low validity score if there are limitation to their external validity.

The review team should carefully document the process followed to assess external validity and to combine the internal / external validity scores in the review protocol, and in the final evidence synthesis report. Each external validity judgement should be accompanied by a concise written justification to reduce subjectivity of interpretation. If any changes to the methods are required during the appraisal process, these should be clearly documented in the final review report as deviations from the protocol. In such cases the updated methods must be applied to all studies included in the review.

f) **Using the results of the critical appraisal.** The results of the critical appraisal should be used to inform the data synthesis (see Step 7, page 14). This applies to both quantitative synthesis (meta-analysis) and qualitative / narrative (descriptive) synthesis. If risks of bias or low external validity are identified in any of the resources included in the review, then the consequence for the data synthesis should be explored and clearly documented, so that the implications for the review's conclusions and recommendations are clear. As for the appraisal of individual studies, numeric scores should not be used for summarising the validity of the overall body of evidence reviewed (for the same reasons outlined under Step 6d above, page 10).

Measures to speed up the appraisal of the evidence for rapid reviews (all should be documented and justified):

- Using a risk of bias tool.
- Limiting risk of bias ratings only to certain form of bias depending on the outcomes of interest for end users and stakeholders.

[12] https://environmentalevidence.org/information-for-authors/7-critical-appraisal-of-study-validity

- Using only one reviewer, but a sample of resources should be examined independently by two reviewers to test for consistency.

7. Data synthesis

Data synthesis refers to the collation of all relevant evidence identified in the review to answer the review question. A review should always have a narrative synthesis of the data, including a tabulation of key characteristics and outcomes of all the resources examined. For Systematic Reviews, if sufficient data is available in a suitable format, a quantitative synthesis, in the form of a meta-analysis, may also be planned.

As for all stages of a review, the data synthesis should follow methods pre-specified in the review protocol, it should be peer-reviewed within the review team and accurately described in the final synthesis report.

a) **Narrative synthesis (both for systematic reviews and systematic maps).** The narrative synthesis presents the context and an overview of the evidence. It includes the tabulation and / or visualisation (often with descriptive statistics) of the findings of the individual studies examined as part of the review with supporting text providing additional information. A narrative synthesis may be the only option for a body of evidence with low validity, or when quantitative data synthesis is not feasible (for example, because there is insufficient quantitative data, or the studies are too dissimilar to be pooled into a meta-analysis). A narrative synthesis is also always present alongside a quantitative synthesis to provide context and background.

   For each study, the following key information should be provided:

   - Study reference
   - Subject population
   - Nature of the impact / exposure
   - Setting / context
   - Outcome measures
   - Methodological design
   - Results of the study, including details of the measured effects (for systematic reviews)
   - Results of the critical appraisal of the study (for systematic reviews)

   The interpretation of the results provided by the authors of the study is not included to ensure that the evidence is presented as objectively as possible.

   Vote counting (e.g., comparing how many studies showed a positive versus negative or neutral outcome based on statistical significance of the results) should be avoided as a form of synthesis. Vote counting is misleading because this procedure does not consider differences in study validity and power.

b) **Quantitative synthesis (for systematic reviews).** A quantitative data synthesis estimates the overall mean and variance of the effect of an impact / exposure by weighting and aggregating the individual effect estimates from all individual studies included in the analysis. With a quantitative data synthesis, it is also possible to investigate sources of heterogeneity (e.g., due diverse environmental conditions) in the contributing studies.

Meta-analysis and meta-regression are the most commonly used methods of quantitative data synthesis in environmental sciences and there is a well-developed supporting literature (Borenstein, 2009; Gates, 2002; Gurevitch & Hedges, 2020; Koricheva, Gurevitch, et al., 2013; Nakagawa et al., 2023) including tutorials and training resources (e.g., https://itchyshin.github.io/Meta-analysis_tutorial/). Details about these methods are also provided by CEE[13]. Therefore, guidance on the use of these methods is not included in this report.

The output of a quantitative data synthesis should include:

- The effect estimate of each study and, where the studies are sufficiently homogeneous, the overall mean effect estimate (i.e., pooled across all studies).

- A forest plot displaying the effect estimates of individual studies as well as the overall effect estimate (if the studies are sufficiently homogenous). An orchard plot can be used as an alternative for large meta-analyses where there are many studies to display.

- Details about the risk of bias for each individual study (i.e., the results of the critical appraisal of each resource). If sufficient data is available, meta-analyses should be undertaken on subgroups (subsets of studies) based on the appraisal of study validity, for example grouping together studies with low, medium and high risk of bias and investigating whether the effect estimate differs between the subgroups. An alternative approach is to run a meta-regression using internal validity (or risk of bias) as a categorical variable.

- An investigation of the heterogeneity of the effect estimates across studies.

It is important to investigate the causes of effect size heterogeneity across studies to properly understand the relevance of the study findings to the review question. A range of sensitivity analyses can be used to explore the factors underlying the effect size variability. Variation in effect sizes across heterogenous studies can be explored by meta-regression. Alternative options could be to exclude the studies identified as outliers and re-run the meta-analysis, or to group together studies with similar characteristics and run separate meta-analyses on those groups of studies. The approach for exploring heterogeneity should be specified *a-priori* in the review protocol.

c) **Mapping and data visualisation (for systematic maps).** The results of systematic maps can be presented in many forms and there are not well-established guidelines. Presentation of the finding of systematic maps can range from simple spreadsheets to innovative forms of data visualisation that can be easily interrogated by the user.

8. Interpretation of the evidence synthesis and reporting

[13] https://environmentalevidence.org/information-for-authors/8-data-synthesis/

Evidence synthesis collates and synthesises data to present reliable evidence in relation to the review question. Authors should simply present the evidence to inform rather than offer advice. When reviews are inconclusive because there is insufficient evidence, it is important not to confuse “no evidence of an effect” (which may indicate the need for further research to build better evidence) with “evidence of no effect” (which instead would suggest that there is enough good-quality evidence to draw this conclusion).

Important aspects to consider when interpreting the findings of a systematic review are:

- The internal / external validity of the resources examined.
- The size and statistical significance of the observed effects.
- The consistency of the effects across studies and the extent to which this can be explained by other variables.
- The clarity of the relationship between the intensity of the impact / exposure and the outcome.
- The existence of any indirect evidence that supports or refutes the inference.
- The existence of bias or confounding effects.

For rapid reviews, the of risk of bias resulting from the modifications to the systematic review methodology should also be discussed.

Guidance on reporting for systematic / rapid reviews and systematic maps are embedded throughout the steps outlined in this section. In addition, CEE has developed standard formats for systematic reviews and maps[14,15] as well as Reporting standards for Systematic Evidence Syntheses (ROSES[16]), which provide a reporting framework for ensuring evidence syntheses report their methods to the highest possible standards.

[14] https://environmentalevidencejournal.biomedcentral.com/submission-guidelines/preparing-your-manuscript/systematic-review

[15] https://environmentalevidencejournal.biomedcentral.com/submission-guidelines/preparing-your-manuscript/systematic-map

[16] https://environmentalevidence.org/roses/

Together. Shaping Better Places.

Boffa Miskell is a leading New Zealand environmental consultancy with nine offices throughout Aotearoa. We work with a wide range of local, international private and public sector clients in the areas of planning, urban design, landscape architecture, landscape planning, ecology, biosecurity, Te Hīhiri (cultural advisory), engagement, transport advisory, climate change, graphics, and mapping. Over the past five decades we have built a reputation for creativity, professionalism, innovation, and excellence by understanding each project's interconnections with the wider environmental, social, cultural, and economic context.

www.boffamiskell.co.nz

Whangarei	Auckland	Hamilton	Tauranga	Wellington	Nelson	Christchurch	Queenstown	Dunedin
09 358 2526	09 358 2526	07 960 0006	07 571 5511	04 385 9315	03 548 8551	03 366 8891	03 441 1670	03 470 0460